Statistica Sinica Preprint No: SS-2025-0104

Title	Efficient Estimation of Average Treatment Effects with Unmeasured Confounding and Proxies
Manuscript ID	SS-2025-0104
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0104
Complete List of Authors	Chunrong Ai and Jiawei Shan
Corresponding Authors	Jiawei Shan
E-mails	jiawei.shan@wisc.edu
Notice: Accepted author version.	

Statistica Sinica

Efficient Estimation of Average Treatment Effects with Unmeasured Confounding and Proxies

Chunrong Ai[1] and Jiawei Shan[2]

[1] *The Chinese University of Hong Kong, Shenzhen*

[2] *University of Wisconsin-Madison*

Abstract: Proximal causal inference provides a framework for estimating the average treatment effect (ATE) in the presence of unmeasured confounding by leveraging outcome and treatment proxies. Identification in this framework relies on the existence of a so-called bridge function. Standard approaches typically postulate a parametric specification for the bridge function, which is estimated in a first step and then plugged into an ATE estimator. However, this sequential procedure suffers from two potential sources of efficiency loss: (i) the difficulty of efficiently estimating a bridge function defined by an integral equation, and (ii) the failure to account for the correlation between the estimation steps. To overcome these limitations, we propose a novel approach that approximates the integral equation with increasing moment restrictions and jointly estimates the bridge function and the ATE. We show that, under suitable conditions, our estimator is efficient. Additionally, we provide a data-driven procedure for selecting the tuning parameter (i.e., the number of moment restrictions). Simulation studies reveal that the proposed method performs well in finite samples, and an application to the right heart catheterization dataset from the SUPPORT study demonstrates its practical value.

Key words and phrases: Data-driven method; generalized method of moments; proximal causal inference; semiparametric efficiency.

1. Introduction

The unconfoundedness assumption is a key condition for identifying treatment effect parameters and establishing the consistency of many popular estimators. This condition may not hold if some confounders are unmeasured and omitted from the empirical analysis. One approach to address the omitted variables problem is proximal causal inference, which assumes the availability of outcome and treatment confounding proxies (Miao et al., 2018; Tchetgen Tchetgen et al., 2024). Such an approach has seen a wide range of applications (Shi et al., 2020; Kallus et al., 2022; Dukes et al., 2023; Ying et al., 2023; Egami and Tchetgen Tchetgen, 2024; Ghassami et al., 2024; Qi et al., 2024; Qiu et al., 2024; Ying, 2024).

With the outcome and treatment confounding proxies, Miao et al. (2018) showed that an *outcome bridge function*, analogous to the outcome regression function one would use if all confounders were observed, identifies the average treatment effect. A common approach is to parameterize the bridge function, estimate it using standard methods, and then estimate the average treatment effect (ATE) via a plug-in. For instance, Miao et al. (2024) suggested a recursive generalized method of moments (RGMM) by transforming the integral equation to fixed-dimensional, user-specified moment restrictions. Tchetgen Tchetgen et al. (2024) introduced a proximal g-computation method, which results in a simple proximal two-stage least squares procedure in the special case of linear working models. The proximal g-computation method is typically more efficient than RGMM, benefiting from an additional parametric restriction on the joint distribution of covariates, but is also more prone to bias if the distribution model is misspecified. These methods suffer from efficiency losses because estimating the bridge function may be inefficient, and the sequential procedure may fail to utilize all information (Brown and Newey, 1998; Ai and Chen, 2012). As a remedy, Cui et al. (2023) proposed a doubly robust (DR) locally efficient approach that incorporates an additional treatment bridge function. They proved that their proximal DR estimator achieves the semiparametric local efficiency bound if both bridge functions are correctly specified and consistently estimated, even if they are not efficiently estimated. If one of the bridge functions is misspecified, the DR estimator fails to achieve local efficiency, and it could benefit from more efficient estimation of the bridge functions. Moreover, the DR estimator may not be the most efficient, even if both bridge functions are consistently estimated, as illustrated in Figure 1.

This paper's primary contribution is to develop a simple, data-driven method for efficiently estimating the average treatment effect. The key step is to transform the conditional moment restriction that defines the bridge function into an expanding set of unconditional moment restrictions via a sieve basis, and to estimate the bridge function and the ATE jointly. As the number of moments increases, these unconditional moment restrictions provide a good approximation for the unknown conditional moment restriction. We show that: (i) the proposed estimator for the bridge function achieves the semiparametric efficiency bound established in Cui et al. (2023), without requiring any modeling of the data distribution beyond the bridge function; (ii) the proposed estimator for ATE has an asymptotic variance that is never larger than that of the DR locally efficient estimator; and (iii) the proposed estimator outperforms the theoretically optimal plug-in estimator. Moreover, the proposed method is readily implemented using widely available software for GMM estimation (Hansen, 1982). We also propose a data-driven procedure for selecting the number of moments.

The theoretical results of this paper are closely related to established theories in causal inference under the assumption of no unmeasured confounding and in the missing data literature under the missing at random assumption, which demonstrate that regression-based estimators are more efficient than doubly robust estimators when the outcome model is correctly specified (Scharfstein et al., 1999; Bang and Robins, 2005). We extend this principle to the proximal causal inference. Our work is also related to several strands of the existing literature. First, it connects to research on instrumental variable estimation and partial identification strategies (Ai and Chen, 2003; Newey and Powell, 2003; Abadie, 2003; Kline and Tamer, 2023). Second, it is related to recent advances that incorporate high-dimensional and machine learning methods into proximal estimation frameworks (Mastouri et al., 2021; Kompa et al., 2022). Third, it is linked to developments in causal graphical models for identifying effects in the presence of unmeasured confounding, including work on the front-door criterion (Pearl, 1995; Richardson et al., 2023; Guo et al., 2023; Bhattacharya et al., 2022; Guo and Nabi, 2024), which also leverage proxy-like mediators to achieve nonparametric identification under structural assumptions.

The rest of the paper is organized as follows. Section 2 reviews the identification results of proximal causal inference and challenges in constructing an efficient estimator. Section 3 describes the proposed method for efficient estimation. Section 4 establishes large-sample properties of the proposed estimators and compares them with the existing methods. Section 5 discusses the choice of tuning parameters for practical implementation. Section 6 provides a simulation study and applies our method to reanalyze the SUPPORT dataset. Section 7 offers a brief discussion. Technical proofs and additional results are provided in the supplementary material.

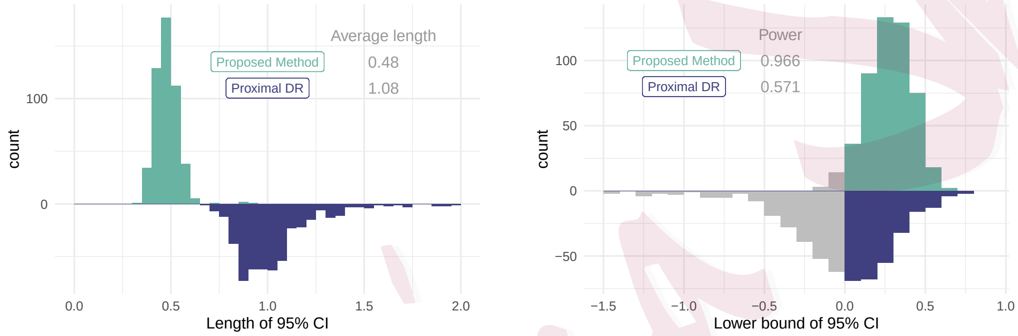


Figure 1: Comparison of two methods with a sample size of 400 over 500 replications. Data is generated following Scenario II in Section 6.1, with a true causal effect of 0.5. The gray shading in the right panel indicates replicates that fail to reject the null hypothesis of no causal effect.

2. Basic framework

Let $A \in \{0, 1\}$ denote the treatment variable, and Y denote the observed outcome. Let $Y(a)$ denote the potential outcome if treatment $A = a$ is assigned. The observed outcome is $Y = Y(A)$. Our objective is to estimate the average treatment effect, $\tau_0 \triangleq \mathbb{E}\{Y(1) - Y(0)\}$, in settings with unmeasured confounders. Let U denote the unmeasured confounders that may affect both the treatment A and outcome Y. Suppose there are three types of observable covariates (X, W, Z): X affect both A and Y; W are outcome-inducing confounding proxies which are related to A only through (X, U); and Z are treatment-inducing confounding proxies which are associated with Y only through (X, U). See Figure 2 for a causal directed acyclic graph (DAG). We formalize the relationships among these variables in the following assumption.

Assumption 1. *We assume that:*

(i) (Latent exchangeability) $Y(a) \perp\!\!\!\perp A \mid (U, X)$ *for* $a = 0, 1$*;*

(ii) (Overlap) $0 < c_1 < \mathrm{pr}(A = a \mid U, X) < c_2 < 1$ *for some constants* c_1 *and* c_2*;*

(iii) (Proxy variables) $Z \perp\!\!\!\perp Y \mid A, U, X$*, and* $W \perp\!\!\!\perp (A, Z) \mid U, X$*;*

(iv) (Completeness) For any square-integrable function g *and any* a, x*,* $\mathbb{E}\{g(U) \mid Z, A = a, X = x\} = 0$ *almost surely if and only if* $g(U) = 0$ *almost surely.*

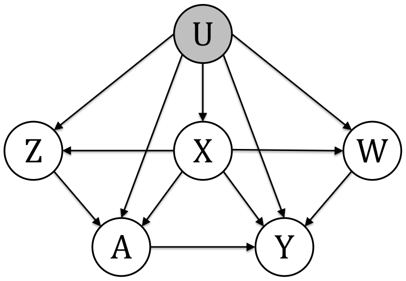


Figure 2: A causal DAG of proximal causal inference with unmeasured confounding and proxies.

Assumption 1(i)-(ii) are standard in the literature of causal inference. Assumption 1(iii) describes the nature of the two types of proxies: treatment proxies Z cannot affect the outcome Y, and the outcome proxies W cannot be affected by either the treatment A or the treatment proxies Z, upon conditioning on the confounders X and U. These conditions are critical for identification but untestable as they involve conditional independence statements given the unmeasured variable U. Justifying these conditions and selecting proxy variables requires subject-matter knowledge; however, they can be satisfied in applications through careful study design. For instance, by incorporating post-outcome variables in Z and pre-treatment measurements of the outcome in W, we can reasonably expect Assumption 1(iii) to hold as the future cannot affect the past. Miao et al. (2024) provided an example of evaluating the short-term effect of air pollution on elderly hospitalization using time series data, in which air pollution measurements taken after hospitalization are included in Z, and hospitalization measurements taken before air pollution are included in W. Assumption 1(iv) can be intuitively interpreted as a requirement for Z to exhibit sufficient variability relative to that of the unmeasured confounder U. In the case of categorical Z and U, it requires Z with at least as many categories as U. In the continuous case, Chen et al. (2014) and Andrews (2017) showed that if the dimension of Z exceeds that of U, then under mild conditions, completeness holds generically, except for distributions in negligible sets. Tchetgen Tchetgen et al. (2024) also suggest measuring a rich set of baseline characteristics to make the completeness assumption more plausible; however, this benefit should be balanced against the efficiency loss incurred by including irrelevant proxies and the increased sample size required for estimating high-dimensional nuisance functions. Completeness is also shown to hold in many parametric and semiparametric models, such as exponential families (Newey and Powell, 2003) and location-scale families (Hu and Shiu, 2018). For a detailed discussion of completeness, see D'Haultfoeuille (2011). We also refer readers to Tchetgen Tchetgen et al. (2024) for a detailed discussion of these assumptions, additional examples of proxies, and approaches for sensitivity analysis in practice.

Under Assumption 1, Miao et al. (2018) established the identification of ATE through an *outcome-confounding bridge function* $h(w, a, x)$, which solves

$$\mathbb{E}(Y \mid Z, A, X) = \mathbb{E}\left\{h(W, A, X) \mid Z, A, X\right\}. \tag{2.1}$$

It is worth noting that the bridge function solves a Fredholm integral equation of the first kind, which is known to be an ill-posed inverse problem (Kress, 1989). Then, ATE is identified by

$$\tau_0 = \mathbb{E}\left\{h(W, 1, X) - h(W, 0, X)\right\}. \tag{2.2}$$

Equation (2.2) is referred to as the *proximal g-formula* in Tchetgen Tchetgen et al. (2024).

We consider the estimation of τ_0 using N independent and identically distributed observations drawn from the joint distribution of $\boldsymbol{O} = (A, Y, W, X, Z)$. We follow Tchetgen Tchetgen et al. (2024) by parameterizing the bridge function $h(W, A, X; \boldsymbol{\gamma})$ with the parameter value $\boldsymbol{\gamma}_0 \in \Gamma \subset \mathbb{R}^p$. Then, one can estimate $\boldsymbol{\gamma}_0$ by solving

$$\frac{1}{N}\sum_{i=1}^{N}\left\{Y_i - h(W_i, A_i, X_i; \boldsymbol{\gamma})\right\}\boldsymbol{m}(Z_i, A_i, X_i) = \mathbf{0}, \tag{2.3}$$

where $\boldsymbol{m}(\cdot)$ is a user-specified vector-valued function, with dimension equal to that of $\boldsymbol{\gamma}$, and satisfies that $\mathbb{E}\left\{\nabla_{\gamma} h(W, A, X; \boldsymbol{\gamma}_0)\boldsymbol{m}(Z, A, X)^{\mathrm{T}}\right\}$ is nonsingular. For instance, if $h(W, A, X; \boldsymbol{\gamma}) = (1, W, A, X)\boldsymbol{\gamma}$, one can choose

$\boldsymbol{m}(Z, A, X) = (1, Z, A, X)^{\mathrm{T}}$. Under mild regularity conditions (White, 1982), $\widehat{\boldsymbol{\gamma}}(\boldsymbol{m})$ that solves (2.3) is consistent and asymptotically normally distributed for any $\boldsymbol{m}$. However, the choice of $\boldsymbol{m}$ may affect the efficiency, namely, the asymptotic variance of $\widehat{\boldsymbol{\gamma}}(\boldsymbol{m})$. Let $\nabla_{\boldsymbol{x}}\boldsymbol{\ell}(\boldsymbol{x}) = \partial\boldsymbol{\ell}(\boldsymbol{x})^{\mathrm{T}}/\partial\boldsymbol{x}$ denote the gradient of a (vector-valued) function $\boldsymbol{\ell}$. Cui et al. (2023) showed that all score functions of $\boldsymbol{\gamma}$, the efficient score is the one with

$$\boldsymbol{m}_{\mathit{eff}}(Z, A, X) = \frac{\mathbb{E}\{\nabla_{\boldsymbol{\gamma}} h(W, A, X; \boldsymbol{\gamma}_0) \mid Z, A, X\}}{\mathbb{E}[\{Y - h(W, A, X)\}^2 \mid Z, A, X]},$$

that accounts for the potential heteroskedasticity. Unfortunately, such $\boldsymbol{m}_{\mathit{eff}}$ is impractical because it requires modeling complex features of the observed data distribution, which are difficult to capture accurately. Miao et al. (2024) suggested a recursive GMM approach, allowing the dimension of $\boldsymbol{m}$ to exceed that of $\boldsymbol{\gamma}$. While their approach may offer some efficiency gains for $\widehat{\boldsymbol{\gamma}}(\boldsymbol{m})$, the specific low-dimensional $\boldsymbol{m}$ does not necessarily lead to an efficient estimator for $\boldsymbol{\gamma}$, in contrast to the increasing moment conditions proposed in this paper, as elaborated later. An inefficient $\widehat{\boldsymbol{\gamma}}(\boldsymbol{m})$ may lead to inefficiency of the plug-in estimator:

$$\widehat{\tau}^{\text{plug-in}}(\boldsymbol{m}) = \frac{1}{N}\sum_{i=1}^{N}\left\{h(W_i, 1, X_i; \widehat{\boldsymbol{\gamma}}(\boldsymbol{m})) - h(W_i, 0, X_i; \widehat{\boldsymbol{\gamma}}(\boldsymbol{m}))\right\}.$$

Furthermore, the plug-in estimators, including $\widehat{\tau}^{\text{plug-in}}(\boldsymbol{m}_{\mathit{eff}})$, fail to account for the correlation between the score functions of $\boldsymbol{\gamma}_0$ and τ_0, thereby introducing an additional source of efficiency loss (Brown and Newey, 1998; Ai and Chen, 2012).

3. Estimation

To improve existing approaches, we propose estimating $\boldsymbol{\gamma}_0$ and τ_0 jointly and using an increasing number of moment restrictions. Specifically, let $\boldsymbol{u}_K(z,a,x) = \big(u_{K1}(z,a,x),\ldots,u_{KK}(z,a,x)\big)^{\mathrm{T}}$ denote a vector of known basis functions (such as power series, splines, Fourier series, etc.) with dimension $K \in \mathbb{N}$, which provides approximation sieves that can approximate a large class of smooth functions arbitrarily well as $K \to \infty$. The selection of $\boldsymbol{u}_K(z,a,x)$ is discussed in Section 5. Model (2.1) implies the following unconditional moment restrictions of $\boldsymbol{\gamma}_0$:

$$\mathbb{E}\left[\{Y - h(W,A,X;\boldsymbol{\gamma}_0)\}\,\boldsymbol{u}_K(Z,A,X)\right] = \mathbf{0}. \tag{3.4}$$

Denote the joint score function

$$\boldsymbol{g}_K(\boldsymbol{O};\boldsymbol{\gamma},\tau) = \begin{pmatrix} \{Y - h(W,A,X;\boldsymbol{\gamma})\}\boldsymbol{u}_K(Z,A,X) \\ \tau - h(W,1,X;\boldsymbol{\gamma}) + h(W,0,X;\boldsymbol{\gamma}) \end{pmatrix},$$

and $\boldsymbol{G}_K(\boldsymbol{\gamma},\tau) = N^{-1}\sum_{i=1}^N \boldsymbol{g}_K(\boldsymbol{O}_i;\boldsymbol{\gamma},\tau)$. Since K increases with the sample size, the number of moment restrictions typically exceeds that of unknown parameters. So, we apply the GMM method for estimation. For a user-specified $(K+1)\times(K+1)$ positive definite matrix $\boldsymbol{\Omega}$, the GMM estimator of $(\boldsymbol{\gamma},\tau)$ is given by

$$(\check{\boldsymbol{\gamma}},\check{\tau}) = \underset{(\boldsymbol{\gamma},\tau)\in\Gamma\times\mathcal{T}}{\arg\min}\ \boldsymbol{G}_K(\boldsymbol{\gamma},\tau)^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{G}_K(\boldsymbol{\gamma},\tau). \tag{3.5}$$

Hansen (1982) showed that, with a fixed $K\geq p$, under some regularity conditions, $(\check{\boldsymbol{\gamma}},\check{\tau})$ are consistent and asymptotically normally distributed but perform best only when $\boldsymbol{\Omega}$ is selected as the inverse of $\boldsymbol{\Upsilon}_{(K+1)\times(K+1)} = \mathbb{E}\{\boldsymbol{g}_K(\boldsymbol{O};\boldsymbol{\gamma}_0,\tau_0)\boldsymbol{g}_K(\boldsymbol{O};\boldsymbol{\gamma}_0,\tau_0)^{\mathrm{T}}\}$. We use the initial estimator $(\check{\boldsymbol{\gamma}},\check{\tau})$ to obtain an estimator of $\boldsymbol{\Upsilon}_{(K+1)\times(K+1)}$:

$$\widehat{\boldsymbol{\Upsilon}}_{(K+1)\times(K+1)} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}_K(\boldsymbol{O}_i;\check{\boldsymbol{\gamma}},\check{\tau})\boldsymbol{g}_K(\boldsymbol{O}_i;\check{\boldsymbol{\gamma}},\check{\tau})^{\mathrm{T}}.$$

We then obtain the optimal GMM estimator

$$(\widehat{\boldsymbol{\gamma}},\widehat{\tau}) = \underset{(\boldsymbol{\gamma},\tau)\in\Gamma\times\mathcal{T}}{\arg\min}\ \boldsymbol{G}_K(\boldsymbol{\gamma},\tau)^{\mathrm{T}}\widehat{\boldsymbol{\Upsilon}}^{-1}_{(K+1)\times(K+1)}\boldsymbol{G}_K(\boldsymbol{\gamma},\tau). \tag{3.6}$$

With a fixed K, Hansen (1982) showed that under regularity conditions,

$$\boldsymbol{V}_K^{-1/2}\begin{pmatrix}\sqrt{N}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\\ \sqrt{N}(\widehat{\tau}-\tau_0)\end{pmatrix}\xrightarrow{d} N\left(\boldsymbol{0},\boldsymbol{I}_{(p+1)\times(p+1)}\right), \tag{3.7}$$

where $\boldsymbol{V}_K = \{\boldsymbol{B}^{\mathrm{T}}_{(K+1)\times(p+1)}\boldsymbol{\Upsilon}^{-1}_{(K+1)\times(K+1)}\boldsymbol{B}_{(K+1)\times(p+1)}\}^{-1}$ and

$$\begin{aligned}\boldsymbol{B}_{(K+1)\times(p+1)} &\triangleq \mathbb{E}\left\{\nabla_{\boldsymbol{\gamma},\tau}\boldsymbol{G}_K(\boldsymbol{\gamma}_0,\tau_0)^{\mathrm{T}}\right\}\\ &= \begin{pmatrix}-\mathbb{E}\{\boldsymbol{u}_K(Z,A,X)\nabla_{\boldsymbol{\gamma}}h(W,A,X;\boldsymbol{\gamma}_0)^{\mathrm{T}}\} & \boldsymbol{0}\\ -\mathbb{E}\left\{\frac{(-1)^{1-A}}{f(A|W,X)}\nabla_{\boldsymbol{\gamma}}h(W,A,X;\boldsymbol{\gamma}_0)^{\mathrm{T}}\right\} & 1\end{pmatrix}.\end{aligned}$$

Remark 1. When K is fixed, the optimal GMM estimator in (3.6) is generally not semiparametrically efficient. In our method, we allow K to increase slowly with the sample size, such that $\{\boldsymbol{u}_K(z,a,x)\}$ spans the space of measurable functions and provides a good approximation of $\boldsymbol{m}_{\mathit{eff}}(z,a,x)$, thereby achieving full efficiency.

Remark 2. Numerical instability may arise if the combined moment conditions (3.4) exhibit near-linear dependence, causing the asymptotic covariance matrix to be singular. When this occurs, some linear combinations of the estimating equations contain negligible independent information about the data-generating process. Therefore, discarding these combinations incurs minimal information loss. Specifically, we recommend a two-step regularization procedure. First, we orthonormalize the basis functions $\boldsymbol{u}_K$ so that their empirical second-moment matrix is the identity, i.e., $N^{-1}\sum_{i=1}^{N}\boldsymbol{u}_K(Z_i,A_i,X_i)\boldsymbol{u}_K(Z_i,A_i,X_i)^{\mathrm{T}}=\boldsymbol{I}_K$. Second, we apply GMM to all estimating equations using the identity weighting (i.e., setting $\boldsymbol{\Omega}=\boldsymbol{I}_{K+1}$ in (3.5)) to obtain initial estimators $(\check{\boldsymbol{\gamma}},\check{\tau})$. We apply spectral decomposition to the estimated moment covariance matrix $\widehat{\boldsymbol{\Upsilon}}_{(K+1)\times(K+1)}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}$. With $\boldsymbol{Q}_{K_1}$ as eigenvectors corresponding to the K_1 eigenvalues larger than a small threshold c, we apply GMM to the linear combination of the estimating equations $\boldsymbol{Q}_{K_1}^{\mathrm{T}}\boldsymbol{g}_K$, using the optimal weight matrix $\boldsymbol{\Lambda}_{K_1}^{-1}$. Here, $\boldsymbol{\Lambda}_{K_1}$

is the diagonal matrix composed of the first K_1 eigenvalues. This procedure effectively filters out highly irrelevant moment conditions and ensures that the weighting matrix is invertible.

Remark 3. The plug-in estimator $\widehat{\tau}^{\text{plug-in}}$ can be interpreted as a GMM estimator that utilizes a block-diagonal weighting matrix. From this perspective, the efficiency loss arises from the suboptimality of this weighting scheme, which fails to exploit the covariance among the estimation equations. In a broader context, this phenomenon mirrors the distinction between simple two-stage estimators (e.g., 2SLS) and joint estimation methods such as efficient GMM or limited-information maximum likelihood (LIML), in which joint estimation achieves greater efficiency by accounting for the correlation structure across stages.

4. Large-sample properties

In this section, we establish the large-sample properties of the proposed estimator as the number of moment restrictions increases. We impose the following assumptions.

Assumption 2. *(i) The eigenvalues of* $\mathbb{E}\{\boldsymbol{u}_K(Z,A,X)\boldsymbol{u}_K(Z,A,X)^{\mathrm{T}}\}$ *are bounded and bounded away from zero for all* K*; (ii) For any p-smooth function* $h(z,a,x)$ *defined in Appendix A, there is* $\boldsymbol{\beta}_h \in \mathbb{R}^K$ *such that*

$\sup_{(z,a,x)\in\mathcal{Z}\times\{0,1\}\times\mathcal{X}} |h(z,a,x) - \boldsymbol{u}_K(z,a,x)^{\mathrm{T}}\boldsymbol{\beta}_h| = O(K^{-\alpha})$ *with* $\alpha > 0$*;*

(iii) $\zeta(K)^2K/N = o(1)$*, where* $\zeta(K) \triangleq \sup_{(z,a,x)\in\mathcal{Z}\times\{0,1\}\times\mathcal{X}} \|\boldsymbol{u}_K(z,a,x)\|$.

Assumption 2(i) rules out the degeneracy of moment restrictions. As indicated in Remark 2, we recommend orthogonalizing the basis functions such that the empirical second-moment matrix is the identity. Assumption 2(ii) requires sieve approximation error rates for the p-smooth function class, which have been well studied in the mathematical literature on approximation theory. For instance, suppose $\mathcal{X}$ and $\mathcal{Z}$ are compact subsets in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_z}$, respectively, and $\boldsymbol{u}_K$ is a tensor product of polynomials or B-splines as introduced in Section 5, Chen (2007, p. 5537) showed that Assumption 2(ii) holds with $\alpha = p/(d_x+d_z)$. Assumption 2(iii) restricts the number of moments to ensure the convergence and asymptotic normality of the proposed estimator. Newey (1997) showed that if $\boldsymbol{u}_K$ is a power series, then $\zeta(K) = O(K)$, and it requires $K = o(N^{1/3})$. If $\boldsymbol{u}_K$ is a B-spline, then $\zeta(K) = O(\sqrt{K})$ and $K = o(\sqrt{N})$. These conditions guide the choice of $\boldsymbol{u}_K$, which is discussed in Section 5. The asymptotic distribution of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\tau}$ is formally established in the following theorem.

Theorem 1. *Suppose Assumptions 1 and 2, and regularity conditions in Appendix A hold. We have* $\sqrt{N}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{V}_\gamma)$ *and* $\sqrt{N}(\widehat{\tau} - \tau_0) \xrightarrow{d}$

$N(0, V_\tau)$ *with* $\boldsymbol{V}_\gamma = \mathbb{E}\{\boldsymbol{\psi}_1(\boldsymbol{O})\boldsymbol{\psi}_1(\boldsymbol{O})^{\mathrm{T}}\}$, $V_\tau = \mathbb{E}\{\psi_2(\boldsymbol{O})^2\}$, *and*

$$\boldsymbol{\psi}_1(\boldsymbol{O}) = \left[\mathbb{E}\left\{\frac{\partial h(W, A, X; \boldsymbol{\gamma}_0)\boldsymbol{m}_{\mathit{eff}}(Z, A, X)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\right\}\right]^{-1}\{Y - h(W, A, X)\}\boldsymbol{m}_{\mathit{eff}}(Z, A, X),$$

$$\psi_2(\boldsymbol{O}) = h(W, 1, X) - h(W, 0, X) - \tau_0 + t(Z, A, X)\{Y - h(W, A, X)\},$$

$$t(Z, A, X) = \boldsymbol{\kappa}^{\mathrm{T}}\boldsymbol{m}_{\mathit{eff}}(Z, A, X) - R(Z, A, X),$$

$$R(Z, A, X) = \frac{\mathbb{E}[\{Y - h(W, A, X)\}\{h(W, 1, X) - h(W, 0, X) - \tau_0\} \mid Z, A, X]}{\mathbb{E}[\{Y - h(W, A, X)\}^2 \mid Z, A, X]},$$

$$\boldsymbol{\kappa} = \boldsymbol{V}_\gamma \mathbb{E}\left[\left\{R(Z, A, X) + \frac{(-1)^{1-A}}{f(A \mid W, X)}\right\}\nabla_\gamma h(W, A, X; \boldsymbol{\gamma}_0)\right].$$

Theorem 1 demonstrates that $(\widehat{\boldsymbol{\gamma}}, \widehat{\tau})$ remains $\sqrt{N}$-consistent and asymptotically normally distributed when the number of moments increases slowly with the sample size. It follows intermediate lemmas in Section S2 of the supplementary material. Specifically, we show that the initial estimators $(\check{\boldsymbol{\gamma}}, \check{\tau})$ obtained from (3.5) are $\sqrt{N}$-consistent. We then demonstrate that (3.7) still holds as $K \to \infty$ slowly. As a result, we obtain Theorem 1 by calculating the limit of $\boldsymbol{V}_K$ in (3.7). Notably, $\boldsymbol{V}_\gamma$ is the semiparametric efficiency bound of $\boldsymbol{\gamma}_0$, derived in Theorem E.2 of Cui et al. (2023). Thus, the proposed estimator of $\boldsymbol{\gamma}_0$ is efficient. To see the efficiency of $\widehat{\tau}$, we consider the semiparametric local efficiency bound of τ_0 derived in Cui et al. (2023, Theorem 3.1), which requires the following assumption.

Assumption 3. *We assume that:*

(i) For any square-integrable function g and for any a, x, $\mathbb{E}\{g(U) \mid W, A =$

$a, X = x\} = 0$ *almost surely if and only if* $g(U) = 0$ *almost surely.*

(ii) There exists a treatment-confounding bridge function $q(z, a, x)$ *that satisfies*

$$\mathbb{E}\{q(Z, A, X) \mid W, A, X\} = \frac{1}{f(A \mid W, X)}. \tag{4.8}$$

(iii) The conditional expectation mappings $T(g) \equiv E\{g(W, A, X) \mid Z, A, X\}$ *and* $T'(g) \equiv E\{g(Z, A, X) \mid W, A, X\}$ *are surjective.*

Assumption 3(i)-(ii) permits an alternative identification formula of ATE, given by $\tau = \mathbb{E}\{(-1)^{1-A} q(Z, A, X) Y\}$, using the treatment-confounding bridge function q instead of h in the proximal g-formula (2.2) as established in Cui et al. (2023). Assumption 3(iii) ensures that h and q are uniquely identified by the integral equations (2.1) and (4.8), respectively. Combining these two identification results, Cui et al. (2023) showed that the efficient influence function of τ_0, under the semiparametric model $\mathcal{M}_{sp}$, which does not restrict the observed data distribution other than the existence of a bridge function h that solves (2.1), evaluated at the submodel where Assumption 3 holds, is

$$\psi_{\mathit{eff}}(\boldsymbol{O}) = h(W, 1, X) - h(W, 0, X) - \tau_0 + (-1)^{1-A} q(Z, A, X)\{Y - h(W, A, X)\}. \tag{4.9}$$

Therefore, the semiparametric local efficiency bound of τ_0 under $\mathcal{M}_{sp}$ is $V_{\tau,\mathit{eff}} = \mathbb{E}\{\psi_{\mathit{eff}}(\boldsymbol{O})^2\}$.

Theorem 2. *(i)* $V_\tau \leq V_{\tau,\mathit{eff}}$*, and the equality holds if and only if there is a vector of constants* $\boldsymbol{\alpha}$ *such that* $R(Z,A,X) + (-1)^{1-A}q(Z,A,X) = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{m}_{\mathit{eff}}(Z,A,X)$*; (ii)* V_τ *is the semiparametric local efficiency bound of* τ_0 *under the submodel* $\mathcal{M}_{sub}$ *where* $h(w,a,x;\boldsymbol{\gamma})$ *is correctly specified and uniquely determined by* (2.1).

Theorem 2 shows that the proposed estimator $\widehat{\tau}$ has an asymptotic variance that is always no larger than the semiparametric efficiency bound $V_{\tau,\mathit{eff}}$. The efficiency gain arises from parameterizing the bridge function h. We note that the doubly robust estimator, constructed using the efficient score (4.9), attains the semiparametric efficiency bound $V_{\tau,\mathit{eff}}$ only if both bridge functions h and q are correctly specified and consistently estimated. We also note that the condition that guarantees $V_\tau = V_{\tau,\mathit{eff}}$ is unlikely to hold, except under highly contrived data distributions. As demonstrated in the simulation studies in Section 6, it does not hold even in the most straightforward linear data-generating processes. Therefore, the proposed estimator generally outperforms the doubly robust estimator equipped with consistently estimated h and q.

Remark 4. Theorems 1 and 2 are established under asymptotics with increasing K. Intuitively, this involves a bias-variance trade-off as highlighted by Donald et al. (2009). A larger K implies using a richer set of moment conditions, which improves estimation efficiency (reduces variance) but can also increase bias. Assumption 2 is crucial here as it restricts K from growing too rapidly, thereby balancing the need for efficiency against the risk of excessive bias. Furthermore, we address the practical selection of K in Section 5 using a data-driven approach based on MSE minimization.

To highlight the deficiency of the plug-in procedure, the following theorem compares the proposed estimator $\widehat{\tau}$ with the theoretically optimal plug-in estimator $\widehat{\tau}^{\text{plug-in}}(\boldsymbol{m}_{\mathit{eff}})$.

Theorem 3. *Suppose that $\sqrt{N}\{\widehat{\tau}^{\text{plug-in}}(\boldsymbol{m}_{\mathit{eff}}) - \tau_0\} \xrightarrow{d} N(0, V_{\tau,\mathit{eff}}^{\text{plug-in}})$, then we have $V_\tau \leq V_{\tau,\mathit{eff}}^{\text{plug-in}}$, and the equality holds if and only if there is a vector of constants $\boldsymbol{\alpha}$ such that $R(Z, A, X) = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{m}_{\mathit{eff}}(Z, A, X)$.*

The condition that guarantees $V_\tau = V_{\tau,\mathit{eff}}^{\text{plug-in}}$ is satisfied if the observed distribution concerning the correlation between the score functions of τ_0 and $\boldsymbol{\gamma}_0$, characterized by $R(Z, A, X)$, meets a particular structure. A sufficient condition is that there is no additive interaction between A and W in the model of $h(W, A, X)$. In this case, equality holds for $\boldsymbol{\alpha} = \mathbf{0}$. In case of its violation, Theorem 3 indicates that the proposed estimator of τ_0

outperforms plug-in estimators, even though an efficient estimator of $\boldsymbol{\gamma}_0$ is employed.

5. Implementation details

The GMM algorithm can be easily implemented using routine software, such as gmm in R. Here, we briefly discuss the selection of tuning parameters. A large class of sieves $\boldsymbol{u}_K(z,a,x)$ is feasible for implementing the proposed approach. In this paper, we suggest a tensor-product linear sieve basis. To simplify the presentation, we suppose both Z and X are scalar variables. Let $\{\varphi_j(x) : j = 1, \ldots, K_1\}$ denote a sieve basis for $\mathcal{L}_2(\mathcal{X})$, the space of square Lebesgue integrable functions on $\mathcal{X}$. For instance, if $\mathcal{X} = [0,1]$, one can choose the basis of polynomials $\varphi_j(x) = x^{j-1}$ or B-spline, and if $\mathcal{X}$ is unbounded, one can select the Hermite polynomial basis $\varphi_j(x) = \exp(-x^2)x^{j-1}$. Similarly, let $\{\phi_k(z) : k = 1, \ldots, K_2\}$ denote a basis for $\mathcal{L}_2(\mathcal{Z})$. Then, the tensor-product sieve basis for $\mathcal{Z} \times \{0,1\} \times \mathcal{X}$ is given by $\{I(a = l)\varphi_j(x)\phi_k(z) : j = 1, \ldots, K_1; k = 1, \ldots, K_2; l = 0, 1\}$ with the number of terms $K = 2K_1K_2$. However, the number of tensor-product basis terms grows exponentially with the dimension of the arguments. In practice, additive separable bases can be employed to mitigate the issue of a large K. We refer readers to Newey (1997) and Chen (2007) for further details on sieve selection. Another tuning parameter is the number of moment restrictions K. While the large-sample properties of the proposed estimator permit a wide range of values for K, practical guidance on selecting smoothing parameters is necessary for applied researchers who generally have only one finite sample at their disposal. Here, we propose an asymptotic mean-square-error-based criterion, as in Donald et al. (2009), for the data-driven selection K. Specifically, we choose $K \in \{1, \ldots, \bar{K}\}$ to minimize $S_{\mathrm{GMM}}(K) := \sum_{j=1}^{p}\{\widehat{\Pi}(K;\boldsymbol{e}_j)^2/N + \widehat{\Phi}(K;\boldsymbol{e}_j)\}$ defined in Appendix B, where $\widehat{\Pi}(K;\boldsymbol{e}_j)^2/N$ is an estimate of the squared bias term of $\widehat{\gamma}_j$ derived in Newey and Smith (2004), and $\widehat{\Phi}(K;\boldsymbol{e}_j)$ is an estimate of the asymptotic variance term of $\widehat{\gamma}_j$. We refer interested readers to Donald et al. (2009) for further insights.

6. Numerical studies

6.1 Simulation

We construct Monte Carlo simulations to examine the finite-sample performance of the proposed approach. We generate i.i.d samples from the following data-generating process:

$$
\begin{aligned}
&X, U \sim N(0,1), \quad \varepsilon_j \mid X = x \sim N\left(0, \sigma_j^2(x)\right), j = 1,2,3, \\
&\operatorname{logit}\{\operatorname{pr}(A = 1 \mid X, U)\} = (1, X, U)\boldsymbol{\beta}_a, \quad Z = (1, A, X, U)\boldsymbol{\beta}_z + \varepsilon_1, \\
&W = (1, X, U)\boldsymbol{\beta}_w + \varepsilon_2, \quad Y = (1, A, W, X, U)\boldsymbol{\beta}_y + \varepsilon_3.
\end{aligned}
$$

with the parameters $\boldsymbol{\beta}_a = (-0.1, 0.5, 0.5)^{\mathrm{T}}$, $\boldsymbol{\beta}_z = (0.5, 1, 0.5, 1)^{\mathrm{T}}$, $\boldsymbol{\beta}_w = (1, -1, 1)^{\mathrm{T}}$, and $\boldsymbol{\beta}_y = (1, 0.5, 0.5, 1, 1)^{\mathrm{T}}$. We consider two scenarios for $\sigma_j(x)$. Scenario I is a simple setting with no heteroskedasticity, $\sigma_j(x) \equiv 1$ for $j = 1, 2, 3$. In Scenario II, heteroskedasticity is present, with $\sigma_1(x) \equiv 1$, $\sigma_2(x) = (0.3 + x^2)^{-1/2}$, $\sigma_3(x) = (0.5 + 0.8x^2)^{-1/2}$. The average treatment effect $\tau_0 = 0.5$. We show in Section S4 of the supplementary material that the above data-generating mechanism is compatible with the following models of h and q:

$$
\begin{aligned}
&h(w, a, x; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 w + \gamma_2 a + \gamma_3 x, \\
&q(z, a, x; \boldsymbol{\theta}) = 1 + \exp\left\{(-1)^a(\theta_0 + \theta_1 z + \theta_2 a + \theta_3 x)\right\}.
\end{aligned}
$$

Our proposed method, denoted GMM-div, is computed using a power series and a data-driven smoothing parameter K with $\bar{K} = 12$ in both scenarios. For comparison, we include five additional methods: (i) Naive estimation, which regards all three covariates (W, Z, X) as confounders, and estimates ATE using classical g-formula (Greenland and Robins, 1986).

(ii) RGMM, recursive GMM estimation with a fixed number of moments introduced in Miao et al. (2024). Following their suggestion, we choose the moment restrictions $\mathbb{E}[\{Y-h(W,A,X;\boldsymbol{\gamma})\}(1,Z,A,X)^{\mathrm{T}}]=0$. In this case, it is equivalent to the *proximal outcome regression* estimator introduced in Cui et al. (2023). (iii) P2SLS, the *proximal two-stage least squares* introduced in Tchetgen Tchetgen et al. (2024), which can be implemented using the command ivreg($Y\sim A+X+W \mid A+X+Z$) in R. (iv) PIPW and PDR, the *proximal inverse probability weighting* and the *proximal doubly robust* estimator introduced by Cui et al. (2023), which require modeling the treatment confounding bridge function $q(z,a,x;\boldsymbol{\theta})$. Specifically, $\widehat{\tau}_{PIPW}=N^{-1}\sum_{i=1}^{N}(-1)^{1-A_i}q(Z_i,A_i,X_i;\widehat{\boldsymbol{\theta}})Y_i$, and $\widehat{\tau}_{DR}$ solves the empirical analogue of (4.9) by replacing h and q with their estimates. Here, $\widehat{\boldsymbol{\theta}}$ is obtained by solving the estimating equation $\mathbb{E}\{(-1)^{1-A}q(Z,A,X;\boldsymbol{\theta})(1,W,A,X)^{\mathrm{T}}-(0,0,1,0)^{\mathrm{T}}\}=0$. Across all methods, we replicate 500 simulations at sample sizes of 400 and 800.

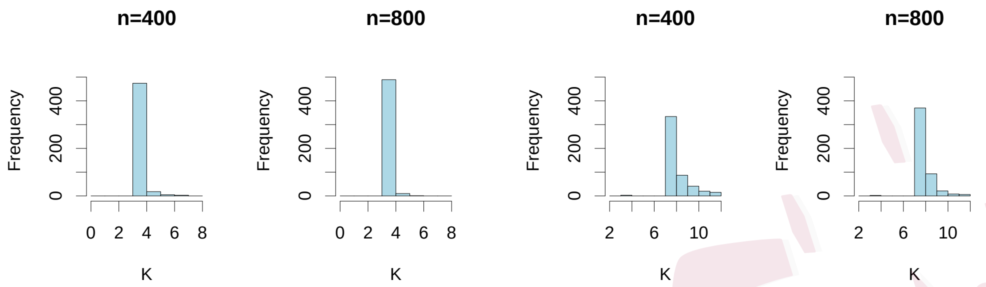
Table 1 reports the absolute bias, standard error, root mean squared error, coverage probability, average length of 95% confidence intervals, and the power of the hypothesis testing problem $H_0:\tau_0=0$ *v.s.* $H_1:\tau_0\neq 0$ in both scenarios. The power is calculated as the proportion of rejected cases across 500 replications. It reveals that: (i) The naive estimator is severely biased due to unmeasured confounding, while the other five methods exhibit negligible bias as expected in both scenarios. (ii) In Scenario I without heteroskedasticity, the five methods perform similarly, and the proposed method mainly selects $K = 4$ across replications as shown in Figure 3(a). (iii) In Scenario II, the proposed method demonstrates significantly lower standard error, narrower confidence intervals, and, notably, higher power in detecting non-zero causal effects. It benefits from the increased efficiency of additional selected moments as shown in Figure 3(b). Figure 4 compares the empirical distributions of different methods, illustrating that the proposed method performs more concentrated around the true value. (iv) Due to the tradeoff between type I and type II errors, the proposed estimator provides a relatively anti-conservative confidence interval. Given the small sample size in Scenario II, this yields a slightly smaller CP than other methods. Nevertheless, it approaches the nominal level as the sample size increases.

6.2 Real data application

We apply the proposed method to re-analyze the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT), as considered in Cui et al. (2023) and Tchetgen Tchetgen et al. (2024). The study aims to evaluate the effectiveness of right heart catheterization

Table 1: Simulation results: absolute bias, standard error (SE), root mean squared error (RMSE), coverage probability (CP), average length of the 95% confidence interval, and power.

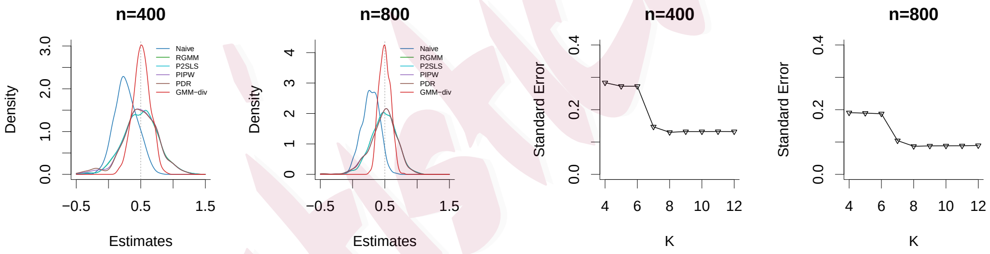
	$n = 400$						$n = 800$					
Method	Bias	SE	RMSE	Length	CP	Power	Bias	SE	RMSE	Length	CP	Power
Scenario I: homoskedasticity												
Naive	0.17	0.13	0.21	0.53	0.776	0.708	0.17	0.10	0.19	0.37	0.580	0.922
RGMM	0.01	0.16	0.16	0.60	0.942	0.888	0.01	0.11	0.11	0.43	0.950	0.988
P2SLS	0.00	0.16	0.16	0.61	0.944	0.898	0.01	0.11	0.11	0.43	0.956	0.986
PIPW	0.00	0.16	0.16	0.64	0.948	0.852	0.01	0.12	0.12	0.45	0.958	0.980
PDR	0.00	0.16	0.16	0.63	0.944	0.854	0.01	0.12	0.12	0.44	0.962	0.982
GMM-div	0.00	0.16	0.16	0.60	0.942	0.886	0.01	0.11	0.11	0.43	0.950	0.988
Scenario II: heteroskedasticity												
Naive	0.23	0.18	0.29	0.71	0.764	0.312	0.22	0.13	0.26	0.50	0.598	0.602
RGMM	0.00	0.28	0.28	1.04	0.954	0.500	0.01	0.19	0.19	0.72	0.952	0.726
P2SLS	0.01	0.28	0.28	1.03	0.952	0.490	0.01	0.19	0.19	0.72	0.948	0.730
PIPW	0.01	0.29	0.29	1.08	0.956	0.534	0.03	0.22	0.22	0.78	0.954	0.734
PDR	0.02	0.31	0.31	1.08	0.950	0.518	0.03	0.23	0.23	0.75	0.942	0.728
GMM-div	0.01	0.13	0.13	0.48	0.936	0.966	0.00	0.09	0.09	0.34	0.956	1.000



(a) Scenario I: homoskedasticity

(b) Scenario II: heteroskedasticity

Figure 3: Histogram of selected K.



(a)

(b)

Figure 4: (a) Empirical distributions of different methods under Scenario II. (b) Standard error of the proposed estimator versus K under Scenario II.

(RHC) in the initial care of critically ill patients (Connors et al., 1996). This dataset has also been widely analyzed in the causal inference literature, assuming no unmeasured confounding (Hirano and Imbens, 2001; Tan, 2006; Vermeulen and Vansteelandt, 2015). The treatment A indicates whether a patient received an RHC within 24 hours of admission. Among 5735 patients, 2184 received the treatment, while 3551 did not. The outcome Y is the number of days between admission and death or censoring at 30 days. The data include 71 baseline covariates, comprising 21 continuous variables and 50 dummy variables derived from categorical variables. These covariates include demographics (age, sex, race, education, income, and insurance status), estimated probability of survival, comorbidity, vital signs, physiological status, and functional status. Following Tchetgen Tchetgen et al. (2024), we select $Z = (\mathsf{pafi1}, \mathsf{paco21})$, $W = (\mathsf{ph1}, \mathsf{hema1})$, and use the remaining covariates as X.

To implement the proposed method, we select $(1, A, Z, X)$, along with quadratic and cubic polynomials of the continuous variables in X, as candidates for constructing moment equations. We then apply the proposed data-driven approach to select K. Figure 5 plots the loss curve against K, from which we choose $K = 81$ for estimation. Our method yields a point estimate of -1.610 with a standard error of 0.272, and the corresponding

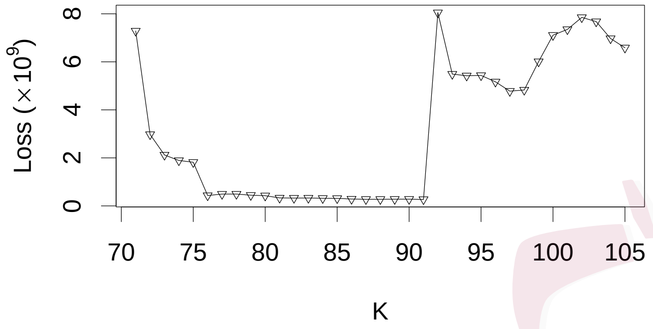


Figure 5: Loss curve for the selection of K in the application of SUPPORT.

95% confidence interval is (-2.143,-1.077). As a comparison, OLS yields an estimate of -1.249 (SE = 0.275) with a 95% CI of (-1.789,-0.709), while the proximal 2SLS proposed by Tchetgen Tchetgen et al. (2024) produces an estimate of -1.798 (SE = 0.431) with a 95% CI of (-2.643,-0.954). Our estimate is closely aligned with the proximal 2SLS estimate, suggesting that OLS, which assumes no unmeasured confounding, may underestimate the harmful effect of RHC on 30-day survival among critically ill patients. Moreover, our estimate has a smaller standard error than the proximal 2SLS estimate, resulting in a narrower confidence interval.

7. Discussion

This paper proposes a simple, data-driven method for obtaining fully efficient estimates of the bridge function and average treatment effect. A key feature of this method is that it requires model assumptions only for the bridge function without imposing any additional assumptions on the data distribution. The proposed estimator typically outperforms the locally efficient DR estimator, demonstrating substantial advantages in detecting significant causal effects. Its feasibility and ease of implementation make it highly useful in practical applications.

Although our method is sufficiently efficient, we acknowledge that it lacks the protective capacity against potential model misspecification that the DR estimator provides. Both reviewers suggested jointly estimating h and q efficiently, similar to our approach, using the DR estimating equation for the treatment effect. However, as noted by Cui et al. (2023), this would not improve the efficiency of the treatment effect estimator. The DR estimating equation is Neyman orthogonal and thus locally insensitive to perturbations in h and q, so efficiency gains in estimating h and q do not translate into efficiency gains for the ATE estimator. This reflects a trade-off between efficiency and robustness. In practice, we therefore recommend using both methods and comparing their results. For example, when the proposed method and the DR method yield similar estimates, it is reasonable to assume the bridge function is correctly specified and to adopt the proposed method's narrower confidence intervals.

Supplementary Materials

The supplementary material includes intermediate lemmas, additional simulation studies, and all technical proofs.

Acknowledgements

Chunrong Ai gratefully acknowledges funding from Project 72133005, supported by the NSFC.

A. Regularity conditions

Smoothness classes of functions: Let $\underline{p}$ be the largest integer satisfying $\underline{p} < p$, and $\boldsymbol{a} = (a_1, \ldots, a_d)$. A function $\phi(\boldsymbol{v})$ with domain $\mathcal{V} \subset \mathbb{R}^d$ is called a *p-smooth function* if it is $\underline{p}$ times continuously differentiable on $\mathcal{V}$ and

$$\max_{\sum_{i=1}^{d} a_i = \underline{p}} \left| \frac{\partial^{\underline{p}} \phi(\boldsymbol{v})}{\partial v_1^{\alpha_1} \cdots \partial v_d^{\alpha_d}} - \frac{\partial^{\underline{p}} \phi(\boldsymbol{v}')}{\partial v_1^{\alpha_1} \cdots \partial v_d^{\alpha_d}} \right| \le C \left\| \boldsymbol{v} - \boldsymbol{v}' \right\|^{p - \underline{p}},$$

for all $\boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}$ and some constant $C > 0$.

Regularity conditions: The following regularity conditions (A1)–(A4) are common in the GMM literature (Newey and McFadden, 1994, p. 2132).

Condition (A5) imposes mild smoothness restrictions to ensure a good sieve approximation.

(A1) Γ is a compact subset in $\mathbb{R}^p$; $\boldsymbol{\gamma}_0$ lies in the interior of Γ and is the unique solution to (3.4); $\mathcal{T}$ is a compact subset in $\mathbb{R}$, and τ_0 lies in the interior of $\mathcal{T}$;

(A2) $h(W, A, X; \boldsymbol{\gamma})$ is twice continuously differentiable in $\boldsymbol{\gamma} \in \Gamma$;

(A3) $\mathbb{E}[\sup_{\boldsymbol{\gamma}\in\Gamma}\{Y - h(W, A, X; \boldsymbol{\gamma})\}^2] < \infty$ and $\mathbb{E}[\sup_{\boldsymbol{\gamma}\in\Gamma} \|\nabla_{\boldsymbol{\gamma}} h(W, A, X; \boldsymbol{\gamma})\|^2] < \infty$;

(A4) $\mathbb{E}[\sup_{\boldsymbol{\gamma}\in\Gamma} \{h(W, 1, X; \boldsymbol{\gamma}) - h(W, 0, X; \boldsymbol{\gamma})\}^2] < \infty$ and $\mathbb{E}[\sup_{\boldsymbol{\gamma}\in\Gamma} \|\nabla_{\boldsymbol{\gamma}}\{h(W, 1, X; \boldsymbol{\gamma}) - h(W, 0, X; \boldsymbol{\gamma})\}\|^2] < \infty$.

(A5) $\mathbb{E}[\nabla_{\boldsymbol{\gamma}} h(W, A, X; \boldsymbol{\gamma}_0) \mid z, a, x]$, $\mathbb{E}[\{Y - h(W, A, X)\}^2 \mid z, a, x]$ and $\mathbb{E}[\{Y - h(W, A, X)\}\{h(W, 1, X) - h(W, 0, X) - \tau_0\} \mid z, a, x]$ are p-smooth functions for some $p > 0$.

B. Notation

The following notations, adapted slightly from Donald et al. (2009), are used in Section 5 for selecting K:

$$\widehat{\boldsymbol{\Upsilon}}_{K\times K} = \frac{1}{N}\sum_{i=1}^{N}\{Y_i - h(W_i, A_i, X_i; \check{\gamma})\}^2 \boldsymbol{u}_K(Z_i, A_i, X_i)\boldsymbol{u}_K(Z_i, A_i, X_i)^{\mathrm{T}},$$

$$\widehat{\boldsymbol{B}}_{K\times p} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}_K(Z_i, A_i, X_i)\nabla_{\boldsymbol{\gamma}} h(W_i, A_i, X_i; \check{\boldsymbol{\gamma}})^{\mathrm{T}}, \quad \widehat{\boldsymbol{\Omega}}_{p\times p} = (\widehat{\boldsymbol{B}}_{K\times p})^{\mathrm{T}}\widehat{\boldsymbol{\Upsilon}}^{-1}_{K\times K}\widehat{\boldsymbol{B}}_{K\times p},$$

$$\widetilde{\boldsymbol{d}}_i = (\widehat{\boldsymbol{B}}_{K\times p})^{\mathrm{T}}\left\{\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{u}_K(Z_j, A_j, X_j)\boldsymbol{u}_K(Z_j, A_j, X_j)^{\mathrm{T}}\right\}^{-1}\boldsymbol{u}_K(Z_i, A_i, X_i),$$

$$\widetilde{\boldsymbol{\eta}}_i = -\nabla_{\boldsymbol{\gamma}} h(W_i, A_i, X_i; \check{\gamma}) - \widetilde{\boldsymbol{d}}_i, \quad \boldsymbol{D}^*_i = (\widehat{\boldsymbol{B}}_{K\times p})^{\mathrm{T}}\widehat{\boldsymbol{\Upsilon}}^{-1}_{K\times K}\boldsymbol{u}_K(Z_i, A_i, X_i),$$

$$\widehat{\xi}_{ij} = \boldsymbol{u}_K(Z_i, A_i, X_i)^{\mathrm{T}}\widehat{\boldsymbol{\Upsilon}}^{-1}_{K\times K}\boldsymbol{u}_K(Z_j, A_j, X_j)/N.$$

For a fixed $\boldsymbol{t} \in \mathbb{R}^p$,

$$\widehat{\Pi}(K; \boldsymbol{t}) = \sum_{i=1}^{N}\widehat{\xi}_{ii}\{Y_i - h(W_i, A_i, X_i; \check{\boldsymbol{\gamma}})\}\boldsymbol{t}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}_{p\times p}\widetilde{\boldsymbol{\eta}}_i,$$

$$\widehat{\Phi}(K; \boldsymbol{t}) = \sum_{i=1}^{N}\widehat{\xi}_{ii}\left\{\boldsymbol{t}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}_{p\times p}\left[\widehat{\boldsymbol{D}}^*_i\{Y_i - h(W_i, A_i, X_i; \check{\boldsymbol{\gamma}})\}^2 + \nabla_{\gamma} h(Z_i, A_i, X_i; \check{\gamma})\right]\right\}^2 - \boldsymbol{t}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}_{p\times p}\boldsymbol{t}.$$

The loss function is

$$S_{\mathrm{GMM}}(K) = \sum_{j=1}^{p}\widehat{\Pi}(K; \boldsymbol{e}_j)^2/N + \widehat{\Phi}(K; \boldsymbol{e}_j),$$

where $\boldsymbol{e}_j$ is the unit vector with 1 in the j-th component and 0 in all others.

In addition, Tables A1–A3 summarize the random variables, notation, and assumptions used in the paper for ease of reference.

Table A1: List of random variables.

A	Binary treatment assignment.
$Y(a)$	Potential outcomes under treatment $A = a$.
Y	Observed outcome.
X	Observed confounders.
U	Unobserved confounders.
(Z, W)	Treatment/outcome-inducing confounding proxies.
$\boldsymbol{O}$	Observed variables $\boldsymbol{O} = (A, Y, W, X, Z)$.

Table A2: List of notation.

$h(w, a, x)$	Outcome-confounding bridge function; see Eq. (2.1).
$q(z, a, x)$	Treatment-confounding bridge function; see Eq. (4.8).
$\boldsymbol{u}_K(z, a, x)$	Vector of basis functions; see Eq. (3.4).
$\boldsymbol{m}_{\mathit{eff}}(z, a, x)$	Efficient score for estimating h; see Eq. (2.3).
$\boldsymbol{g}_K(\boldsymbol{O}; \gamma, \tau)$	Joint score function for estimating h and τ_0; see Section 3.
$\boldsymbol{G}_K(\gamma, \tau)$	Sample average of $\boldsymbol{g}_K(\boldsymbol{O}; \gamma, \tau)$; see Section 3.
$\boldsymbol{\Upsilon}_{(K+1)\times(K+1)}$	Optimal weight in GMM estimation; see Section 3.
$\boldsymbol{B}_{(K+1)\times(p+1)}$	Jacobian matrix; see Eq. (3.7).
$\{\boldsymbol{\psi}_1(\boldsymbol{O}), \psi_2(\boldsymbol{O})\}$	Influence functions of $(\widehat{\boldsymbol{\gamma}}, \widehat{\tau})$; see Theorem 1.
$\psi_{\mathit{eff}}(\boldsymbol{O})$	Efficient influence function of τ_0; see Eq. (4.9).
$\{t(\cdot), R(\cdot), \boldsymbol{\kappa}\}$	Terms in the influence function; see Theorem 1.
$\boldsymbol{V}_K$	Asymptotic variance of the GMM estimator; see Eq. (3.7).
$(\boldsymbol{V}_{\boldsymbol{\gamma}}, V_\tau)$	Asymptotic variance of $(\widehat{\boldsymbol{\gamma}}, \widehat{\tau})$; see Theorem 1.
$V_{\tau,\mathit{eff}}$	Semiparametric efficiency bound for τ_0; see Eq. (4.9).
$V_{\tau,\mathit{eff}}^{\text{plug-in}}$	Variance of the optimal plug-in estimator; see Theorem 3.

Table A3: Summary of assumptions.

Assumption 1	Conditions for identifying treatment effects using proxies.
Assumption 2	Conditions for the use of sieve techniques.
Assumption 3	Conditions for developing semiparametric theory.
Conditions (A1)-(A5)	Conditions for the asymptotic analysis of the GMM estimator.

References

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

Ai, C. and X. Chen (2003, November). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Ai, C. and X. Chen (2012, October). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics 170*(2), 442–457.

Andrews, D. W. K. (2017, August). Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics 199*(2), 213–220.

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

Bhattacharya, R., R. Nabi, and I. Shpitser (2022, January). Semiparametric inference for causal effects in graphical models with hidden variables. *J. Mach. Learn. Res. 23*(1), 295:13325–295:13400.

Brown, B. W. and W. K. Newey (1998). Efficient semiparametric estimation of expectations. *Econometrica 66*(2), 453–464.

Chen, X. (2007, January). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, pp. 5549–5632. Elsevier.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014). Local identification of nonparametric and semiparametric models. *Econometrica 82*(2), 785–809.

Connors, Jr, A. F., T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, Jr, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson, Jr, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus (1996, September). The effectiveness of right heart catheterization in the initial care of critically III patients. *JAMA 276*(11), 889–897.

Cui, Y., H. Pu, X. Shi, W. Miao, and E. Tchetgen Tchetgen (2023). Semiparametric proximal causal inference. *Journal of the American Statistical Association 0*(0), 1–12.

D'Haultfoeuille, X. (2011, June). On the completeness condition in nonparametric instrumental problems. *Econometric Theory 27*(3), 460–471.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009, September). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics 152*(1), 28–36.

Dukes, O., I. Shpitser, and E. J. Tchetgen Tchetgen (2023, March). Proximal mediation analysis. *Biometrika*, asad015.

Egami, N. and E. J. Tchetgen Tchetgen (2024, April). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology 86*(2), 487–511.

Ghassami, A., A. Yang, I. Shpitser, and E. Tchetgen Tchetgen (2024, July). Causal inference with hidden mediators. *Biometrika*, asae037.

Greenland, S. and J. M. Robins (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology 15*(3), 413–419.

Guo, A., D. Benkeser, and R. Nabi (2023, December). Targeted machine learning for average causal effect estimation using the front-door functional. *arXiv preprint arXiv:2312.10234*.

Guo, A. and R. Nabi (2024, September). Average causal effect estimation in DAGs with hidden variables: Extensions of back-door and front-door criteria. *arXiv preprint arXiv:2409.03962*.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Hirano, K. and G. W. Imbens (2001, December). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology 2*(3), 259–278.

Hu, Y. and J.-L. Shiu (2018, June). Nonparametric identification using instrumental variables: Sufficient conditions for completeness. *Econometric Theory 34*(3), 659–693.

Kallus, N., X. Mao, and M. Uehara (2022, October). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*.

Kline, B. and E. Tamer (2023, September). Recent developments in partial identification. *Annual Review of Economics 15*(Volume 15, 2023), 125–150.

Kompa, B., D. R. Bellamy, T. Kolokotrones, J. Robins, and A. Beam (2022, October). Deep learning methods for proximal inference via maximum moment restriction. In *Advances in Neural Information Processing Systems*.

Kress, R. (1989). *Linear Integral Equations*, Volume 82 of *Applied Mathematical Sciences*. New York: Springer New York.

Mastouri, A., Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. Kusner, A. Gretton, and K. Muandet (2021, July). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7512–7523. PMLR.

Miao, W., Z. Geng, and E. Tchetgen Tchetgen (2018, December). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika 105*(4), 987–993.

Miao, W., X. Shi, Y. Li, and E. J. Tchetgen Tchetgen (2024, October). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields 8*(4), 262–273.

Newey, W. K. (1997, July). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. and D. McFadden (1994, January). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Volume 4, pp. 2111–2245. Elsevier.

Newey, W. K. and J. L. Powell (2003, September). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K. and R. J. Smith (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica 72*(1), 219–255.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Uncertainty in Artificial Intelligence. Proceedings of the Eleventh Conference (1995)*, 435–43.

Qi, Z., R. Miao, and X. Zhang (2024, April). Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association 119*(546), 915–928.

Qiu, H., X. Shi, W. Miao, E. Dobriban, and E. Tchetgen Tchetgen (2024, June). Doubly robust proximal synthetic controls. *Biometrics 80*(2), ujae055.

Richardson, T. S., R. J. Evans, J. M. Robins, and I. Shpitser (2023, February). Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics 51*(1), 334–361.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999, December). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Shi, X., W. Miao, J. C. Nelson, and E. J. Tchetgen Tchetgen (2020, April). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(2), 521–540.

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

Tchetgen Tchetgen, E. J., A. Ying, Y. Cui, X. Shi, and W. Miao (2024). An introduction to proximal causal inference. *Statistical Science in press*.

Vermeulen, K. and S. Vansteelandt (2015, September). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association 110*(511), 1024–1036.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Ying, A. (2024, May). Proximal survival analysis to handle dependent right censoring. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae037.

Ying, A., W. Miao, X. Shi, and E. J. Tchetgen Tchetgen (2023, July). Proximal causal inference for complex longitudinal studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(3), 684–704.

Chunrong Ai

School of Management and Economics, Chinese University of Hong Kong, Shenzhen

E-mail: chunrongai@cuhk.edu.cn

Jiawei Shan

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

E-mail: jiawei.shan@wisc.edu