Two-Step Estimation and Inference with Possibly Many Included Covariates*
Supplemental Appendix

Matias D. Cattaneo† Michael Jansson‡ Xinwei Ma§

August 30, 2018

Abstract

This Supplemental Appendix contains general theoretical results encompassing those discussed in the main paper, includes the proofs of these general results, discusses additional methodological and technical results, applies the general results to several treatment effect, policy evaluation and applied microeconomics examples, and reports additional details on the empirical applications and simulation study presented in the main paper.

*This paper encompasses and supersedes our previous paper titled "Marginal Treatment Effects with Many Instruments", presented at the 2016 NBER summer meetings. We specially thank Pat Kline for posing a question that this paper answers, and Josh Angrist, Guido Imbens and Ed Vytlacil for very useful comments on an early version of this paper. We also thank the Editor, Aureo de Paula, three anonymous reviewers, Lutz Kilian, Whitney Newey and Chris Taber for very useful comments. The first author gratefully acknowledges financial support from the National Science Foundation (SES 1459931). The second author gratefully acknowledges financial support from the National Science Foundation (SES 1459967) and the research support of CREATES (funded by the Danish National Research Foundation under grant no. DNRF78). Disclaimer: This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here do not necessarily reflect the views of the BLS.

†Department of Economics, Department of Statistics, University of Michigan.

‡Department of Economics, UC Berkeley and *CREATES*.

§Department of Economics, University of Michigan.

Contents

SA-1 Setup and Main Assumptions

This document is self-contained. We employ the same notation as in the main paper, but we reintroduce the setup and assumptions to facilitate cross-referencing herein. Given a random sample $\{\mathbf{w}_i, \mu_i\}_{1\leq i\leq n}$, we are interested in estimating the population parameter $\boldsymbol{\theta}_0$, which is defined by the following moment condition:

$$\mathbb{E}\left[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\right] = \mathbf{0}, \tag{E.1}$$

where $\mathbf{m}$ is a known moment function. Recall that $\{\mu_i\}_{1\leq i\leq n}$ are not directly observed. Instead, the observed data is $\mathbf{w}_i = [\mathbf{y}_i^\mathsf{T}, r_i, \mathbf{z}_i^\mathsf{T}]^\mathsf{T}$, with $r_i \in \mathbb{R}$ and $\mathbf{z}_i \in \mathbb{R}^k$ satisfying the following first step generated regressors condition:

$$\begin{aligned} r_i &= \mu_i \qquad\quad + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i|\mathbf{z}_i] = 0 \\ &= \mathbf{z}_i^\mathsf{T}\boldsymbol{\beta} + \eta_i + \varepsilon_i, \qquad \mathbb{E}[\mathbf{z}_i\eta_i] = \mathbf{0}. \end{aligned} \tag{E.2}$$

The disturbance ε_i can be interpreted as a structural error term, or simply the error of a conditional expectation decomposition. The only substantive restriction is $\mu_i = \mathbb{E}[r_i|\mathbf{z}_i]$, as explained in the main paper. On the other hand, η_i arises without loss of generality because it captures the misspecification error coming from using the best linear approximation to the unknown conditional expectation.

Estimating $\boldsymbol{\theta}_0$ is straightforward via Generalized Method of Moments (GMM), which leads to the following two-step procedure:

$$\hat{\boldsymbol{\theta}}: \left|\frac{1}{n}\boldsymbol{\Omega}_n^{1/2}\sum_{i=1}^n \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\right|^2 \leq \inf_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\boldsymbol{\Omega}_n^{1/2}\sum_{i=1}^n \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta})\right|^2 + o_{\mathbb{P}}(1), \tag{E.3}$$

$$\hat{\mu}_i = \mathbf{z}_i^\mathsf{T}\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_i \left(r_i - \mathbf{z}_i^\mathsf{T}\boldsymbol{\beta}\right)^2, \tag{E.4}$$

where $\Theta \subset \mathbb{R}^{d_\theta}$ is the parameter space and $|\cdot|$ is the Euclidean norm. To derive distributional properties, we will use the first-order condition:

$$\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\right]^\mathsf{T} \boldsymbol{\Omega}_n \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\right] = o_{\mathbb{P}}(1). \tag{E.5}$$

Let C denote a generic nonnegative and finite constant, whose exact definition depends on the specific context. We omit the subscript n whenever possible, and limits are taken with respect to $n \to \infty$, unless otherwise specified. For two non-negative sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n \precsim b_n$ if and only if $a_n \leq C_n \cdot b_n$, where $C_n = O(1)$ if non-random and $C_n = O_{\mathbb{P}}(1)$ if random.

A random variable is said to be in BM_ℓ (bounded moments) if its ℓ-th moment is finite, and in BCM_ℓ (bounded conditional moments) if its ℓ-th conditional (on $\mathbf{z}_i$) moment is bounded uniformly

by a finite constant. Let $\dot{\mathbf{m}}(\cdot) = \partial \mathbf{m}(\cdot)/\partial \mu$, and $\ddot{\mathbf{m}}(\cdot) = \partial^2 \mathbf{m}(\cdot)/\partial \mu^2$. When evaluated at the true parameters, we further write $\mathbf{m}_i = \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$ and analogously for $\dot{\mathbf{m}}_i$ and $\ddot{\mathbf{m}}_i$. We also define the transformation

$$\mathcal{H}_i^{\alpha,\delta}(\mathbf{m}) = \sup_{(|\mu-\mu_i|+|\boldsymbol{\theta}-\boldsymbol{\theta}_0|)^\alpha \leq \delta} \frac{|\mathbf{m}(\mathbf{w}_i, \mu, \boldsymbol{\theta}) - \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|}{(|\mu - \mu_i| + |\boldsymbol{\theta} - \boldsymbol{\theta}_0|)^\alpha}.$$

Equivalently, it is true that $|\mathbf{m}(\mathbf{w}_i, \mu, \boldsymbol{\theta}) - \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)| \leq \mathcal{H}_i^{\alpha,\delta}(\mathbf{m}) \cdot (|\mu - \mu_i| + |\boldsymbol{\theta} - \boldsymbol{\theta}_0|)^\alpha$ in a small neighborhood. The same transformations are also applied to derivatives of $\mathbf{m}$.

Assumption A.1 (Setup)**.** Let $0 < \delta$, α, $C < \infty$ be some fixed constants.

A.1(1) There is a random sample $\{\mathbf{w}_i\}_{1 \leq i \leq n}$ satisfying (E.1) and (E.2), where $\boldsymbol{\theta}_0 \in \Theta$ is the unique and interior root of (E.1).

A.1(2) There exists positive semi-definite weighting matrices $\{\boldsymbol{\Omega}_n\}_{n \geq 1}$, such that the probability limit $\boldsymbol{\Omega}_n \to_{\mathbb{P}} \boldsymbol{\Omega}_0$ is positive definite.

A.1(3) $\hat{\boldsymbol{\theta}}$ satisfies (E.3), (E.5), and is tight.

A.1(4) $\mathcal{H}_i^{\alpha,\delta}(\mathbf{m}) \in \mathsf{BM}_1$.

A.1(5) $\mathbf{m}$ and $\dot{\mathbf{m}}$ are continuously differentiable in $\boldsymbol{\theta}$ with $\mathcal{H}_i^{\alpha,\delta}(\partial \mathbf{m}/\partial \boldsymbol{\theta}), \mathcal{H}_i^{\alpha,\delta}(\partial \dot{\mathbf{m}}/\partial \boldsymbol{\theta}) \in \mathsf{BM}_1$. Further, the matrix $\mathbf{M}_0 = \mathbb{E}\left[\partial \mathbf{m}_i/\partial \boldsymbol{\theta}\right]$ has full (column) rank d_θ.

A.1(6) $\mathbf{m}$ is twice continuously differentiable in μ.

A.1(7) $\mathbf{m}_i$, $\dot{\mathbf{m}}_i$, $\ddot{\mathbf{m}}_i$, $\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}})$, ε_i^3, $|\dot{\mathbf{m}}_i \varepsilon_i|$, $|\ddot{\mathbf{m}}_i|\varepsilon_i^2$, $|\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}})|\varepsilon_i^2 \in \mathsf{BCM}_2$. ⌟

The next set of assumptions impose smoothness and bounded moments on various quantities.

Assumption A.2 (First-Step Covariates)**.**

A.2(1) $\max_{1 \leq i \leq n} |\hat{\mu}_i - \mu_i| = o_{\mathbb{P}}(1)$.

A.2(2) The approximation error η_i in (E.2) satisfies $\mathbb{E}[\eta_i^2] = o(n^{-1/2})$ and $\mathbb{E}[|\boldsymbol{\zeta}_i|^2]\mathbb{E}[\eta_i^2] = o(n^{-1})$, where $\boldsymbol{\zeta}_i = \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i] - \boldsymbol{\Gamma}\mathbf{z}_i$ with the matrix $\boldsymbol{\Gamma}$ such that $\mathbb{E}[\mathbf{z}_i \boldsymbol{\zeta}_i^\mathsf{T}] = \mathbf{0}$. ⌟

SA-2 Primitive Conditions for First-Step Estimation

A key assumption in our paper is the uniform consistency of the first step estimate (A.2(1)). We discuss primitive conditions for this assumption. We have

$$\max_{1 \leq i \leq n} \left|\hat{\mu}_i - \mu_i\right| \leq \max_{1 \leq i \leq n} \left|\eta_i - \sum_j \pi_{ij}\eta_j\right| + \max_{1 \leq i \leq n} \left|\sum_j \pi_{ij}\varepsilon_j\right|, \tag{E.6}$$

where π_{ij} denotes the (i,j) element of the projection matrix $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-}\mathbf{Z}^\mathsf{T}$, with $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]^\mathsf{T}$. We study each term above to show that $\max_{1\leq i\leq n} |\hat{\mu}_i - \mu_i| \to_{\mathbb{P}} 0$, and give in the process easy-to-interpret primitive conditions for specific types of covariates $\mathbf{z}_i$.

SA-2.1 Linear Approximation Error

Using elementary and Cauchy-Schwarz inequalities, we obtain

$$\max_{1\leq i\leq n} \left|\eta_i - \sum_j \pi_{ij}\eta_j\right| \leq \max_{1\leq i\leq n} |\eta_i| \left(1 + \max_{1\leq i\leq n} \sum_j |\pi_{ij}|\right) \leq \max_{1\leq i\leq n} |\eta_i| \left(1 + \sqrt{n \max_{1\leq i\leq n} \sum_j \pi_{ij}^2}\right).$$

Because $\sum_j \pi_{ij}^2 = \pi_{ii}$, the term above will vanish in probability if

$$\max_{1\leq i\leq n} |\eta_i| = o_{\mathbb{P}}\left(\frac{1}{1 + \sqrt{n \max_{1\leq i\leq n} \pi_{ii}}}\right).$$

Therefore, two sufficient conditions to control the first bounding term in (E.6) are: (i) the approximation error is uniformly small $\max_{1\leq i\leq n} |\eta_i| = o_{\mathbb{P}}(1)$, and (ii) $(\max_{1\leq i\leq n} |\eta_i|^2) \cdot (\max_{1\leq i\leq n} \pi_{ii}) = o_{\mathbb{P}}(n^{-1})$. We discuss some methods to control $\max_{1\leq i\leq n} \pi_{ii}$ below. For example, when the covariates $\mathbf{z}_i$ are locally supported basis expansions (e.g., Splines), $\max_{1\leq i\leq n} \pi_{ii} = O_{\mathbb{P}}(k/n)$, under regularity conditions, so that $\max_{1\leq i\leq n} |\eta_i| = o_{\mathbb{P}}(n^{-1/4})$ will suffice.

SA-2.2 Residual Variability

Because $\mathbb{E}[\varepsilon_i|\mathbf{Z}] = 0$, sharp probability bounds can be established for the second bounding term in (E.6). We illustrate the case when ε_i has sub-Gaussian tail and hence an exponential bound can be used, but other assumptions and probability inequalities are possible depending on the primitive assumptions imposed. A general version of the Hoeffding's inequality applied to sub-Gaussian random variables implies (Vershynin, 2018, Theorem 2.6.2)

$$\begin{aligned}\mathbb{P}\left[\left|\max_{1\leq i\leq n} \sum_j \pi_{ij}\varepsilon_j\right| \geq t \,\middle|\, \mathbf{Z}\right] &\leq n \cdot \max_{1\leq i\leq n} \mathbb{P}\left[\left|\sum_j \pi_{ij}\varepsilon_j\right| \geq t \,\middle|\, \mathbf{Z}\right] \leq n \cdot \max_{1\leq i\leq n} 2\exp\left(-\frac{Ct^2}{\sum_j \pi_{ij}^2 M_j^2}\right) \\ &\leq 2\exp\left(-\frac{Ct^2}{(\max_{1\leq i\leq n} \pi_{ii})(\max_{1\leq i\leq n} M_i^2)} + \log(n)\right),\end{aligned}$$

where $M_i = \inf\{t \geq 0 : \mathbb{E}[\exp\{\varepsilon_i^2/t^2\}|\mathbf{z}_i] \leq 2\}$ is the conditional ψ_2-norm of ε_i, and hence

$$\left(\max_{1\leq i\leq n} \pi_{ii}\right) \cdot \left(\max_{1\leq i\leq n} M_i^2\right) = o_{\mathbb{P}}\left(\frac{1}{\log(n)}\right) \qquad \Rightarrow \qquad \mathbb{P}\left[\left|\max_{1\leq i\leq n} \sum_j \pi_{ij}\varepsilon_j\right| \geq t\right] \to 0,$$

for any t. Therefore, as for the first bounding term in (E.6) discussed above, the result follows by properties of the possibly many covariates $\mathbf{z}_i$ through the statistic $\max_{1\leq i\leq n} \pi_{ii}$.

SA-2.3 Bounding $\max_{1\le i\le n}\pi_{ii}$

The results above showed that the properties of the first step estimator can be determined by studying the behavior of the statistic $\max_{1\le i\le n}\pi_{ii}$, which in turn depends on the properties of $\mathbf{z}_i$. We now study these properties, and give concrete examples for specific types of covariates.

Let $\lambda_{\min}(\mathbf{A})$ denote the minimum eigenvalue of a matrix $\mathbf{A}$. Then, it follows immediately that

$$\frac{k}{n} \;\le\; \max_{1\le i\le n}\pi_{ii} \;\le\; \min\left\{\frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)}\frac{\max_{1\le i\le n}|\mathbf{z}_i|^2}{n}\;,\;1\right\}.$$

The upper bound can be used to give primitive conditions on different types of covariates $\mathbf{z}_i$. Here we focus on bounding $\max_{1\le i\le n}|\mathbf{z}_i|^2$ first, and then deducing the restrictions required on $\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)$. We offer several examples showcasing how meaningful bounds can be obtained in various contexts. Of course, the list of examples above is not meant to be exhaustive, nor the bounds given are supposed to be tight. This list nonetheless is useful to illustrate the wide applicability of our results.

Case 1: Locally Bounded Series Expansions

If $\mathbf{z}_i$ is formed using locally bounded series expansions such as splines, compact-supported wavelets or partitioning basis, then

$$\max_{1\le i\le n}\pi_{ii} \precsim_{\mathbb{P}} \frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)}\cdot\frac{k}{n},$$

and $\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)$ is bounded away from zero with probability approaching one, under regularity conditions, provided that $k\log(k)/n\to 0$. See Belloni et al. (2015) and Cattaneo, Farrell and Feng (2018) for more discussion and examples.

Case 2: Bounding through the Orlicz Norm

A more general approach employs the Orlicz norm. Let ψ be a convex, nondecreasing and nonzero function such that $\psi(0)=0$ and $\limsup_{x,y\to\infty}\psi(x)\psi(y)/\psi(cxy)<\infty$ for some constant c, then van der Vaart and Wellner (1996, Lemma 2.2.2) states that

$$\Big\|\max_{1\le i\le n}|\mathbf{z}_i|^2\Big\|_\psi \precsim \psi^{-1}(n)\cdot\||\mathbf{z}_i|^2\|_\psi \le k\cdot\psi^{-1}(n)\cdot\max_{1\le\ell\le k}\|z_{i,\ell}^2\|_\psi,$$

where $\{z_{i,\ell}:1\le\ell\le k\}$ are the elements of $\mathbf{z}_i$, and the norm is defined as $\|x\|_\psi=\inf\{C>0:\mathbb{E}[\psi(|x|/C)]\le 1\}$.

First consider $\psi(x)=|x|^\alpha$ for some $\alpha>2$, then the above reduces to

$$\Big(\mathbb{E}\Big[\max_{1\le i\le n}|\mathbf{z}_i|^{2\alpha}\Big]\Big)^{1/\alpha} \precsim k\cdot n^{1/\alpha}\cdot\max_{1\le\ell\le k}\Big(\mathbb{E}[z_{i,\ell}^{2\alpha}]\Big)^{1/\alpha},$$

which essentially requires bounding higher moments of the coordinates of $\mathbf{z}_i$, and therefore implies

$$\max_{1\leq i\leq n}\pi_{ii} \precsim_{\mathbb{P}} \frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)}\cdot\frac{k}{\sqrt{n}}\frac{1}{n^{1/2-1/\alpha}}\cdot\max_{1\leq\ell\leq k}\left(\mathbb{E}[|z_{i,\ell}|^{2\alpha}]\right)^{1/\alpha} = o_{\mathbb{P}}(1),$$

provided that $n^{1/\alpha-1/2}\max_{1\leq\ell\leq k}\left(\mathbb{E}[|z_{i,\ell}|^{2\alpha}]\right)^{1/\alpha}/\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n) = o_{\mathbb{P}}(1)$.

If, instead, $\mathbf{z}_i$ has sub-Gaussian tail, then we can use $\psi(x)=\exp(|x|)$, which implies that ($|\mathbf{z}_i|^2$ has sub-exponential tail):

$$\max_{1\leq i\leq n}\pi_{ii} \precsim_{\mathbb{P}} \frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)}\cdot\frac{k}{\sqrt{n}}\frac{\log(n)}{\sqrt{n}}\cdot\max_{1\leq\ell\leq k}\|z_{i,\ell}^2\|_\psi = o_{\mathbb{P}}(1),$$

provided that $\log(n)n^{-1/2}\max_{1\leq\ell\leq k}\|z_{i,\ell}^2\|_\psi/\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n) = o_{\mathbb{P}}(1)$.

Case 3: Regression with dummy variables

It is not uncommon to encounter regression specifications including many dummy variables, such as year/region/group specific fixed effects or expansions of categorical/factor variables, and interactions among them. We illustrate how $\max_{1\leq i\leq n}\pi_{ii}$ can be controlled when many dummy variables are included. Let $\{z_{i,\ell}\}_{1\leq\ell\leq k}$ be the coordinates of $\mathbf{z}_i$, with $z_{i,\ell}\in\{0,1\}$ and $\sum_\ell z_{i,\ell}=1$. Despite the fact that $|\mathbf{z}_i|=1$, hence it must be sub-Gaussian, the coordinates are highly correlated and therefore it is very hard to control the ψ_2-norm of the vector. On the other hand, we still have the bound

$$\max_{1\leq i\leq n}\pi_{ii} \leq \frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)}\cdot\frac{|\mathbf{z}_i|^2}{n} = O_{\mathbb{P}}\left(\frac{1}{n}\frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)}\right).$$

Let $N_\ell=\sum_i z_{i,\ell}$ be the number of observations for which the ℓ-th dummy variable takes value 1, and $p_{n,\ell}=\mathbb{P}[z_{i,\ell}=1]$ and $\underline{p}_n=\min_{1\leq\ell\leq k}p_{n,\ell}$. Since, of course, a dummy variable will not be included for a category with zero observations, we assume without loss of generality that $N_\ell>0$. (In practice, this can be justified using a generalized inverse, that is, $(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)^-$.) Therefore,

$$\frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)} = \frac{n}{\min_{1\leq\ell\leq k}\{N_\ell : N_\ell>0\}} \leq \frac{n}{\min_{1\leq\ell\leq k}N_\ell}.$$

It is easy to see that, under the conditions given below, $\mathbb{P}[\min_{1\leq\ell\leq k}N_\ell>0]\to 1$. To see its asymptotic order, apply Bernstein's inequality (for some $t\in(0,1)$):

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{\lambda_{\min}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/n)}\geq\frac{t}{\underline{p}_n}\right] \leq \mathbb{P}\left[\min_{1\leq\ell\leq k}N_\ell\leq t^{-1}n\underline{p}_n\right] \leq \sum_{\ell=1}^k\mathbb{P}\left[N_\ell\leq t^{-1}np_{n,\ell}\right]\\
&=\sum_{\ell=1}^k\mathbb{P}\left[N_\ell-np_{n,\ell}\leq np_{n,\ell}(t^{-1}-1)\right] \leq \sum_{\ell=1}^k\exp\left[-\frac{1}{2}\frac{np_{n,\ell}(t^{-1}-1)^2}{(1-p_{n,\ell})+\frac{1}{3}(t^{-1}-1)}\right]\to 0,
\end{aligned}$$

provided that (i) $\max_{1\le\ell\le k} p_{n,\ell} \to 0$ and (ii) $n \min_{1\le\ell\le k} p_{n,\ell}/\log(k) \to \infty$. Then, we have

$$\frac{1}{\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z}/n)} = O_{\mathbb{P}}\left(\frac{1}{\min_{1\le\ell\le k} p_{n,\ell}}\right).$$

One example is a balanced design, where for two constants C_1 and C_2, $C_1 k^{-1} \le \min_{1\le\ell\le k} p_{n,\ell} \le \max_{1\le\ell\le k} p_{n,\ell} \le C_2 k^{-1}$. Then under the assumption $k = O(\sqrt{n})$,

$$\max_{1\le i\le n} \pi_{ii} = O_{\mathbb{P}}\left(\frac{k}{n}\right) = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).$$

SA-2.4 Design Balance

In linear regression with increasing dimensions, the design matrix plays an important role in determining the properties of the estimated coefficients and the linear predictors. We already encountered one notion of design balance in the previous subsection, namely $\max_{1\le i\le n} \pi_{ii} = O_{\mathbb{P}}(r_n)$ for some $r_n \downarrow 0$. Note that we do not impose this assumption in the paper, and instead we make the high-level uniform consistency assumption A.2(1). The reason is simple: in concrete examples, it might be easier to exploit the specific structure of the covariates to justify the uniform consistency assumption.

There are other concepts of design balance, which we do assume in Section SA-6 to show the validity of the jackknife. Since those conditions are tightly connected to the previous subsection, we give some remarks here, aiming to clarify their connections.

Recall the π_{ij} is the (i,j)-th element of the projection matrix, which is of rank k with probability approaching one, and therefore $\sum_i \pi_{ii} = k$. Intuitively, the "distribution" of π_{ii} should not be too concentrated on any i, and hence $\sum_i \pi_{ii}^2 = o_{\mathbb{P}}(k)$. This is one notion of design balance we employ. Another assumption we make is $\max_{1\le i\le n} 1/(1-\pi_{ii}) = O_{\mathbb{P}}(1)$. Intuitively, this implies that the diagonal elements of the projection matrix do not have probability mass at 1 asymptotically, since otherwise it would not be possible to "delete one" observation and recompute the projection matrix.

It is easy to see that $\max_{1\le i\le n} \pi_{ii} = o_{\mathbb{P}}(1)$ is a stronger notion of design balance, because

$$\max_{1\le i\le n} \pi_{ii} = o_{\mathbb{P}}(1) \quad \Rightarrow \quad \sum_i \pi_{ii}^2 = o_{\mathbb{P}}(k)$$

$$\max_{1\le i\le n} \pi_{ii} = o_{\mathbb{P}}(1) \quad \Rightarrow \quad \max_{1\le i\le n} 1/(1-\pi_{ii}) \to_{\mathbb{P}} 1 = O_{\mathbb{P}}(1).$$

However, an interesting question is whether the converse is also true. In the following example we show that the two weaker notions of design balance can hold, even when $\max_{1\le i\le n} \pi_{ii} \neq o_{\mathbb{P}}(1)$. This example also gives a clear justification of why we do not explicitly assume $\max_{1\le i\le n} \pi_{ii} = o_{\mathbb{P}}(1)$ anywhere in the paper.

Example (Dummies with small cell-probability)**.** Consider the last example introduced in the previous subsection: Regression with dummy variables. We continue to use the same notation

given above, and hence let N_ℓ be the number of observations such that $z_{i,\ell}$ takes value 1 (i.e., number of observations in cell ℓ), and $p_{n,\ell}$ be the cell probability (i.e., $p_{n,\ell} = \mathbb{P}[z_{i,\ell} = 1]$).

However, we now consider the extreme scenario where the first cell probability satisfies $p_{n,1} = c/n$ for some $c > 0$, and for $\ell \geq 2$ the probabilities are $p_{n,\ell} = (1 - c/n)/(k-1)$. This captures the empirically relevant case where some cells may have very few observations. The problem here, however, is that N_1 follows the Binomial$(n, c/n)$ distribution, which is asymptotically Poisson(c) distributed. And by the discussion in previous subsection, we have

$$\max_{1\leq i\leq n} \pi_{ii} = \max\left\{\pi_{11},\ \max_{2\leq i\leq n} \pi_{ii}\right\} \rightsquigarrow \frac{1}{P}\mathbb{1}[P > 0], \qquad P \sim \text{Poisson}(c),$$

since $\max_{2\leq i\leq n} \pi_{ii} = o_{\mathbb{P}}(1)$.

In reality, one will not include a dummy variable if it is only "on" for one or two observations in the sample. Hence, in our current example, the first covariate is added to the regression if and only if $N_1 \geq C$, where $C \geq 2$ is some pre-specified value. Note that this strategy is legitimate in practice because the model selection is done without referring to the outcome variable. In fact, methods involving recursive partitioning or partitioning by quantiles set a lower limit on the cell size, which corresponds to C, and a low cell probability occurs if the density of the underlying variable is close to zero.

Therefore, when the first covariate is only included when $N_1 \geq C$, we have:

$$\max_{1\leq i\leq n} \pi_{ii} = \max\left\{\pi_{11},\ \max_{2\leq i\leq n} \pi_{ii}\right\} \rightsquigarrow \frac{1}{P}\cdot\mathbb{1}[P \geq C], \qquad P \sim \text{Poisson}(c).$$

In this practically relevant case, it follows immediately that $\max_{1\leq i\leq n} \pi_{ii}$ does not vanish, and still $\max_{1\leq i\leq n} 1/(1-\pi_{ii})$ remains bounded in probability. Finally,

$$\begin{aligned}
\sum_i \pi_{ii}^2 &= \sum_i \sum_{\ell=1}^k \left(\frac{1}{N_\ell}\right)^2 \mathbb{1}[N_\ell \geq C]\mathbb{1}[z_{i,\ell} = 1] \\
&= \sum_{\ell=1}^k N_\ell \left(\frac{1}{N_\ell}\right)^2 \mathbb{1}[N_\ell \geq C] = \sum_{\ell=1}^k \frac{1}{N_\ell}\mathbb{1}[N_\ell \geq C] \\
&\leq \frac{1}{N_1}\mathbb{1}[N_1 \geq C] + \sum_{\ell=2}^k \frac{1}{N_\ell} = O_{\mathbb{P}}(1) + o_{\mathbb{P}}(k-1) = o_{\mathbb{P}}(k),
\end{aligned}$$

where the $o_{\mathbb{P}}(k-1)$ term comes from the discussion in the previous subsection. ⌟

SA-3 The Effect of Including Many Covariates

The first result is the consistency of $\hat{\boldsymbol{\theta}}$.

Theorem SA.1 (Consistency).
If A.1(1)–A.1(4) and A.2(1) hold, then $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| = o_{\mathbb{P}}(1)$. ⌟

Next we consider large sample properties of the two-step GMM estimator $\hat{\boldsymbol{\theta}}$. We have

$$o_{\mathbb{P}}(1) = \mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_n \left[\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\right] + \mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_n \left[\frac{1}{n}\sum_i \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\right] \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right),$$

where we assume $\hat{\boldsymbol{\theta}} \to_{\mathbb{P}} \boldsymbol{\theta}_0$, with $\boldsymbol{\theta}_0$ an interior point of Θ, and $\tilde{\boldsymbol{\theta}}$ denotes a linear combination between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.

Lemma SA.2.
If A.1 and A.2 hold, and $k = O(\sqrt{n})$, then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \boldsymbol{\Sigma}_0 \left[\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\right]\Big(1 + o_{\mathbb{P}}(1)\Big), \tag{E.7}$$

where $\boldsymbol{\Sigma}_0 = -(\mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_0\mathbf{M}_0)^{-1}\mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_0$. ⌟

A Taylor expansion with respect to the first-step estimate, $\hat{\mu}_i$, gives

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) && \text{(E.8)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big) && \text{(E.9)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big)^2 && \text{(E.10)}\\
&+ o_{\mathbb{P}}(1).
\end{aligned}$$

The following lemma shows that (E.9) contributes to not only the asymptotic variance, but also the asymptotic bias.

Lemma SA.3.
If A.1 and A.2 hold, and $k = O(\sqrt{n})$, then

$$\text{(E.9)} = \frac{1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right)\cdot\varepsilon_i + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{1,i}\cdot\pi_{ii} + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_{1,i} = \mathbb{E}\Big[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\cdot\varepsilon_i\ \Big|\mathbf{z}_i\Big]$. If, in addition, $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(1)$, then

$$\frac{1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right)\cdot\varepsilon_i = \frac{1}{\sqrt{n}}\sum_i \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i]\cdot\varepsilon_i + o_{\mathbb{P}}(1).$$

⌟

Inspection of the proof of this lemma shows that only $\mathbb{E}[\eta_i^2] = o(1)$ and $\mathbb{E}[|\boldsymbol{\zeta}_i|^2]\mathbb{E}[\eta_i^2] = o(n^{-1})$ is required; the stronger assumption $\mathbb{E}[\eta_i^2] = o(n^{-1/2})$ will be used when studying the quadratic term

(E.10) in the expansion. Furthermore, when $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(1)$, this lemma shows that it is possible to drop the double sum as well as the projection matrix in the variance component, leading to an asymptotic linear representation.

The following lemma shows that the quadratic term (E.10) also contributes a bias.

Lemma SA.4.
If A.1 and A.2 hold, and $k = O(\sqrt{n})$, then

$$(\text{E.10}) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_{2,ij} = \frac{1}{2}\mathbb{E}\Big[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \cdot \varepsilon_j^2 \ \Big| \mathbf{z}_i, \mathbf{z}_j\Big]$. ⌟

The following theorem combines the previous lemmas, and gives the asymptotic representation of the estimator $\hat{\boldsymbol{\theta}}$ when $k = O(\sqrt{n})$

Theorem SA.5 (Asymptotic Representation)**.**
If A.1 and A.2 hold, and $k = O(\sqrt{n})$, then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{\boldsymbol{\mathcal{B}}}{\sqrt{n}}\right) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_{\mathbb{P}}(1),$$

where

$$\boldsymbol{\mathcal{B}} = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \left[\sum_i \mathbf{b}_{1,i}\pi_{ii} + \sum_{i,j} \mathbf{b}_{2,ij}\pi_{ij}^2\right]$$

$$\bar{\boldsymbol{\Psi}}_1 = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \left[\sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\right] \qquad \bar{\boldsymbol{\Psi}}_2 = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \left[\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0) \ |\mathbf{z}_j]\pi_{ij}\right) \cdot \varepsilon_i\right].$$

⌟

In this Supplemental Appendix, we use $\boldsymbol{\mathcal{B}}$ to denote the bias term. Note that $\boldsymbol{\mathcal{B}} = O_{\mathbb{P}}(k/\sqrt{n})$ hence is non-vanishing under the assumption that $k \propto \sqrt{n}$. The term $\boldsymbol{\mathcal{B}}$ can be viewed as the bias of the limiting distribution. In the main paper, we use $\mathscr{B}$ to denote the bias of $\hat{\boldsymbol{\theta}}$. The two terms are connected through the $\sqrt{n}$-scaling: $\boldsymbol{\mathcal{B}} = \sqrt{n}\mathscr{B}$.

In addition, for the asymptotic representation, we use

$$\boldsymbol{\Psi}_i = \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0) \ |\mathbf{z}_j]\pi_{ij}\right) \cdot \varepsilon_i,$$

and therefore $\bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 = \boldsymbol{\Sigma}_0 \sum_i \boldsymbol{\Psi}_i/\sqrt{n}$.

Theorem SA.6 (Asymptotic Normality)**.**
If the assumptions of Theorem SA.5 hold, then

$$\Big(\mathbb{V}[\mathbb{E}[\bar{\boldsymbol{\Psi}}_1|\mathbf{Z}]] + \mathbb{V}[\bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2|\mathbf{Z}]\Big)^{-\frac{1}{2}}\Big(\bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2\Big) \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

provided that $\mathbb{V}[\bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2|\mathbf{Z}]$ has minimum eigenvalue bounded away from zero with probability approaching one. ⌟

SA-4 Extensions

We discuss three extensions of our basic framework: (i) multidimensional first-step estimator, (ii) semi-linear first-step estimator, and (iii) high-dimensional covariates entering the second-step estimating equation.

SA-4.1 First Step: Multidimensional Case

Generalizing to vector-valued $\boldsymbol{\mu}_i$ is an easy and natural extension of our results, although the notation required becomes more delicate/complicated. We show now how the asymptotic representation of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ changes when there are multiple unknowns estimated in the first step. To illustrate, we discuss in more detail the nature of the many covariates bias in the special case of $\boldsymbol{\mu}_i$ bivariate.

The second step estimating equation takes the following form:

$$\mathbf{0} = \mathbb{E}[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)], \qquad \boldsymbol{\mu}_i = \begin{bmatrix}\mu_{1i} & \mu_{2i} & \cdots & \mu_{d_\mu i}\end{bmatrix}^{\mathsf{T}},$$

where the vector of unknowns $\boldsymbol{\mu}_i$ has dimension d_μ and has to be estimated in the first step. The first step takes the same form,

$$r_{\ell i} = \mu_{\ell i} + \varepsilon_{\ell i} = \mathbf{z}_i^{\mathsf{T}}\boldsymbol{\beta}_\ell + \eta_{\ell i} + \varepsilon_{\ell i}, \qquad 1 \leq \ell \leq d_\mu,$$

with $\eta_{\ell i}$ being the approximation error and $\varepsilon_{\ell i}$ being the error from a conditional expectation decomposition, i.e. $\mathbb{E}[\mathbf{z}_i\eta_{\ell i}] = 0$ and $\mathbb{E}[\varepsilon_{\ell i}|\mathbf{z}_i] = 0$ for $1 \leq \ell \leq d_\mu$. We allow for different sets of covariates being used in each first step estimate ($\ell = 1, 2, \ldots, d_\mu$) or, alternatively, think of $\mathbf{z}_i$ as a "long vector" which collects jointly the covariates used for estimating each $\mu_{\ell i}$.

Both notation and assumptions need to be adjusted in this generalized setting. Define

$$\frac{\partial}{\partial \mu_\ell}\mathbf{m}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) = \dot{\mathbf{m}}_\ell(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0), \qquad \frac{\partial^2}{\partial \mu_\ell \partial \mu_{\ell'}}\mathbf{m}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) = \ddot{\mathbf{m}}_{\ell\ell'}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0),$$

where $1 \leq \ell, \ell' \leq d_\mu$. Modified assumptions are postponed to the end of this section. The general-

ized asymptotic expansion is

$$\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big) = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \tag{E.11}$$

$$+ \boldsymbol{\Sigma}_0 \sum_{\ell=1}^{d_\mu} \left[\frac{1}{\sqrt{n}} \sum_i \Big(\sum_j \mathbb{E}[\dot{\mathbf{m}}_\ell(\mathbf{w}_j, \boldsymbol{\mu}_j, \boldsymbol{\theta}_0) | \mathbf{z}_j] \pi_{ij} \Big) \cdot \varepsilon_{\ell i} \right] \tag{E.12}$$

$$+ \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} \tag{E.13}$$

$$+ \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + o_{\mathbb{P}}(1). \tag{E.14}$$

As before, (E.11) represents the influence function had $\boldsymbol{\mu}_i$ been observed, and (E.12) is the variance contribution from estimating $\boldsymbol{\mu}_i$. The bias terms (E.13) and (E.14) are also the natural generalization of our main result:

$$\mathbf{b}_{1,i} = \sum_{\ell=1}^{d_\mu} \mathbb{E}[\dot{\mathbf{m}}_\ell(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \varepsilon_{\ell i} | \mathbf{z}_i] \tag{E.15}$$

$$\mathbf{b}_{2,ij} = \sum_{\ell,\ell'=1}^{d_\mu} \frac{1}{2} \mathbb{E}[\ddot{\mathbf{m}}_{\ell\ell'}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \varepsilon_{\ell j} \varepsilon_{\ell' j} | \mathbf{z}_i, \mathbf{z}_j]. \tag{E.16}$$

Here we use i and j to index observations, and ℓ and ℓ' to index elements in the unknown vector $\boldsymbol{\mu}_i$. We therefore define the following quantities:

$$\boldsymbol{\mathcal{B}} = \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \left[\sum_i \mathbf{b}_{1,i} \pi_{ii} + \mathbf{b}_{2,ij} \pi_{ij}^2 \right],$$

$$\bar{\boldsymbol{\Psi}}_1 = \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \left[\sum_i \mathbf{m}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \right], \qquad \bar{\boldsymbol{\Psi}}_2 = \sum_{\ell=1}^{d_\mu} \left[\frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_i \Big(\sum_j \mathbb{E}[\dot{\mathbf{m}}_\ell(\mathbf{w}_j, \boldsymbol{\mu}_j, \boldsymbol{\theta}_0) | \mathbf{z}_j] \pi_{ij} \Big) \cdot \varepsilon_{\ell i} \right],$$

then the analogue of Theorems SA.5 and Theorem SA.6 hold.

We will not repeat the argument for the jackknife or the bootstrap, since there is no difficulty in generalizing them to vector-valued $\boldsymbol{\mu}_i$. For the bootstrap, however, we make one remark in Section SA-7 to emphasize how the first step is bootstrapped in this setting.

Finally, the following adjustments have to be made to our assumptions:

Assumption (Vector-Valued $\boldsymbol{\mu}_i$)**.**

A.1(6) $\rightarrow$ $\mathbf{m}$ is twice continuously differentiable in $\boldsymbol{\mu}$, with derivatives denoted by $\dot{\mathbf{m}}_\ell$ and $\ddot{\mathbf{m}}_{\ell\ell'}$, respectively.

A.1(7) $\rightarrow$ For all $1 \leq \ell, \ell' \leq d_\mu$, $\mathbf{m}_i$, $\dot{\mathbf{m}}_{\ell,i}$, $\ddot{\mathbf{m}}_{\ell\ell',i}$, $\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}}_{\ell\ell'})$, $|\boldsymbol{\varepsilon}_i|^2$, $|\dot{\mathbf{m}}_{\ell,i}||\boldsymbol{\varepsilon}_i|$, $|\ddot{\mathbf{m}}_{\ell\ell',i}||\boldsymbol{\varepsilon}_i|^2$, $|\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}}_{\ell\ell'})||\boldsymbol{\varepsilon}_i|^2 \in$ BCM_2.

A.2(1) $\to \max_{1\leq i\leq n}|\hat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i| = o_{\mathbb{P}}(1)$.

A.2(2) $\to \mathbb{E}[|\boldsymbol{\eta}_i|^2] = o(1/\sqrt{n})$ and $\mathbb{E}[|\boldsymbol{\zeta}_{\ell i}|^2]\mathbb{E}[\eta_{\ell i}^2] = o(1/n)$, where $\boldsymbol{\zeta}_{\ell i} = \mathbb{E}[\dot{\mathbf{m}}_\ell(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\mathbf{z}_i] - \boldsymbol{\Gamma}_\ell\mathbf{z}_i$ with $\mathbb{E}[\mathbf{z}_i\boldsymbol{\zeta}_{\ell i}^{\mathsf{T}}] = \mathbf{0}$. ⌟

SA-4.1.1 Special Case: Bivariate $\boldsymbol{\mu}_i$

Let $\boldsymbol{\mu}_i = [\mu_{1i},\mu_{2i}]^{\mathsf{T}}$. Starting from the sample estimating equation, and linearizing with respect to $\hat{\boldsymbol{\theta}}$, we obtain (c.f., Lemma SA.2):

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = \boldsymbol{\Sigma}_0\left[\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i,\hat{\boldsymbol{\mu}}_i,\boldsymbol{\theta}_0)\right]\Big(1+o_{\mathbb{P}}(1)\Big),$$

where

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i,\hat{\boldsymbol{\mu}}_i,\boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0) && \text{(E.17)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}_1(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_{1i}-\mu_{1i}\Big) && \text{(E.18)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}_2(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_{2i}-\mu_{2i}\Big) && \text{(E.19)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \frac{1}{2}\ddot{\mathbf{m}}_{11}(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_{1i}-\mu_{1i}\Big)^2 && \text{(E.20)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \frac{1}{2}\ddot{\mathbf{m}}_{22}(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_{2i}-\mu_{2i}\Big)^2 && \text{(E.21)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \ddot{\mathbf{m}}_{12}(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_{1i}-\mu_{1i}\Big)\Big(\hat{\mu}_{2i}-\mu_{2i}\Big) && \text{(E.22)}\\
&+ o_{\mathbb{P}}(1).
\end{aligned}$$

As before, (E.17) is the influence function if $\boldsymbol{\mu}_i$ were observed; (E.18)–(E.19) contribute the linear (leave-in) bias and variance from estimating $\boldsymbol{\mu}_i$; and (E.20)–(E.22) give the quadratic biases.

As in Lemma SA.3, we have the following for (E.18) and (E.19):

$$\begin{aligned}
\text{(E.18)} &= \frac{1}{\sqrt{n}}\sum_i\left(\sum_j \mathbb{E}[\dot{\mathbf{m}}_1(\mathbf{w}_j,\boldsymbol{\mu}_j,\boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij}\right)\cdot\varepsilon_{1i} + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{1,1,i}\pi_{ii} + o_{\mathbb{P}}(1),\\
\text{(E.19)} &= \frac{1}{\sqrt{n}}\sum_i\left(\sum_j \mathbb{E}[\dot{\mathbf{m}}_2(\mathbf{w}_j,\boldsymbol{\mu}_j,\boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij}\right)\cdot\varepsilon_{2i} + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{1,2,i}\pi_{ii} + o_{\mathbb{P}}(1),
\end{aligned}$$

where $\mathbf{b}_{1,1,i} = \mathbb{E}[\dot{\mathbf{m}}_1(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\varepsilon_{1i}|\mathbf{z}_i]$ and $\mathbf{b}_{1,2,i} = \mathbb{E}[\dot{\mathbf{m}}_2(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\varepsilon_{2i}|\mathbf{z}_i]$.

The quadratic terms (E.20) and (E.21) are handled by Lemma SA.4:

$$(\text{E.20}) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,11,ij} \pi_{ij}^2 + o_{\mathbb{P}}(1) \qquad (\text{E.21}) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,22,ij} \pi_{ij}^2 + o_{\mathbb{P}}(1),$$

with $\mathbf{b}_{2,11,ij} = \frac{1}{2}\mathbb{E}[\ddot{\mathbf{m}}_{11}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \cdot \varepsilon_{1j}^2 | \mathbf{z}_i, \mathbf{z}_j]$ and $\mathbf{b}_{2,22,ij} = \frac{1}{2}\mathbb{E}[\ddot{\mathbf{m}}_{22}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \cdot \varepsilon_{2j}^2 | \mathbf{z}_i, \mathbf{z}_j]$.

Finally, for the new cross-term (E.22), the Cauchy-Schwarz inequality implies

$$\begin{aligned}
|(\text{E.22})| &\leq \frac{1}{\sqrt{n}} \sum_i |\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)| \cdot |\hat{\mu}_{1i} - \mu_{1i}| \cdot |\hat{\mu}_{2i} - \mu_{2i}| \\
&\leq \sqrt{\frac{1}{\sqrt{n}} \sum_i |\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)|^2 \cdot |\hat{\mu}_{1i} - \mu_{1i}|^2} \sqrt{\frac{1}{\sqrt{n}} \sum_i |\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)|^2 \cdot |\hat{\mu}_{2i} - \mu_{2i}|^2} \precsim_{\mathbb{P}} \frac{k}{\sqrt{n}},
\end{aligned}$$

but we would like to have a more precise characterization. Following the same strategy to prove Lemma SA.4, we have the following approximation. We focus on the leading term (conditional expectation calculation), and omit the remainder (conditional variance calculation) to conserve space. We have

$$\begin{aligned}
(\text{E.22}) &= \frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \Big(\hat{\mu}_{1i} - \mu_{1i} \Big) \Big(\hat{\mu}_{2i} - \mu_{2i} \Big) \\
&= \frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_{1j} \Big) \Big(\sum_j \pi_{ij} \varepsilon_{2j} \Big) + o_{\mathbb{P}}(1),
\end{aligned}$$

where the extra $o_{\mathbb{P}}(1)$ corresponds to terms involving the approximation errors η_{1i} and η_{2i}. Then,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_{1j} \Big) \Big(\sum_j \pi_{ij} \varepsilon_{2j} \Big) \,\middle|\, \mathbf{Z} \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,j'} \mathbb{E}\Big[\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \pi_{ij} \pi_{ij'} \varepsilon_{1j} \varepsilon_{2j'} \Big| \mathbf{Z} \Big] \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}\Big[\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \pi_{ij} \pi_{ij} \varepsilon_{1j} \varepsilon_{2j} \Big| \mathbf{z}_i, \mathbf{z}_j \Big] = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,12,ij} \cdot \pi_{ij}^2,
\end{aligned}$$

where $\mathbf{b}_{2,12,ij} = \mathbb{E}[\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \varepsilon_{1j} \varepsilon_{2j} | \mathbf{z}_i, \mathbf{z}_j]$, and we ignored terms with $j \neq j'$ from the second to the third line since the conditional expectation is zero. There are some interesting observations regarding this new bias term. First, if the cross derivative has zero conditional mean (i.e. $\mathbb{E}[\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) | \mathbf{z}_i] = 0$), this bias will be of order $\sum_i \pi_{ii}^2 / \sqrt{n}$. For example, when $\mathbf{m}$ is linearly additive in the two unknowns μ_{1i} and μ_{2i}. Second, if correlation between the two error terms is zero (i.e. $\mathbb{E}[\varepsilon_{1j} \varepsilon_{2j} | \mathbf{z}_j] = 0$), the bias contribution from this term is again of order $\sum_i \pi_{ii}^2 / \sqrt{n}$. To give a concrete example, consider the two unknowns being estimated with independent samples in the first step. In Section SA-6, we will assume $\sum_i \pi_{ii}^2 = o_{\mathbb{P}}(k)$ for the validity of the jackknife. Under this additional assumption, the new bias will be negligible if either the cross derivative has

zero conditional mean or the error terms have zero conditional correlation.

SA-4.2 First Step: Partially Linear Case

As a second generalization of our main results, we discuss a first step estimation model taking a partially linear structure. To be more specific, we partition $\mathbf{z}_i \in \mathbb{R}^{d_\gamma+k}$ into $\mathbf{z}_{1i} \in \mathbb{R}^{d_\gamma}$ and $\mathbf{z}_{2i} \in \mathbb{R}^k$, and consider the following first step:

$$r_i = \nu_i + \varepsilon_i = \mathbf{z}_{1i}^\mathsf{T}\boldsymbol{\gamma} + \mu_i + \varepsilon_i = \mathbf{z}_{1i}^\mathsf{T}\boldsymbol{\gamma} + \mathbf{z}_{2i}^\mathsf{T}\boldsymbol{\beta} + \eta_i + \varepsilon_i = \mathbf{z}_i^\mathsf{T}\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\beta}\end{bmatrix} + \eta_i + \varepsilon_i,$$

with the requirement that $\mathbb{E}[\eta_i \mathbf{z}_i] = \mathbf{0}$ and $\mathbb{E}[\varepsilon_i|\mathbf{z}_i] = 0$, so that η_i continues to be the approximation error and ε_i is the residual from the conditional expectation decomposition. The vector $\boldsymbol{\beta}$ has dimension k, which increases with the sample size, while $\boldsymbol{\gamma}$ has a fixed dimension d_γ. A canonical example is $\mu_i = \mu(\tilde{\mathbf{z}}_i)$ being an unknown function and $\mathbf{z}_{2i}$ being series expansion of a collection of covariates.

The second step is also generalized:

$$\mathbb{E}[\mathbf{m}(\mathbf{w}_i,\ \mu_i,\ \boldsymbol{\gamma},\ \boldsymbol{\theta}_0)] = \mathbf{0},$$

where now both $\boldsymbol{\gamma}$ and μ_i enter the second step estimating equation. The real difficulty here is that μ_i *is no longer a conditional expectation projection*, unless $\boldsymbol{\gamma}$ is known or $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are orthogonal. To make progress, we rewrite the problem as as follows:

$$\mathbb{E}[\mathbf{m}(\mathbf{w}_i,\ \nu_i - \mathbf{z}_{1i}^\mathsf{T}\boldsymbol{\gamma},\ \boldsymbol{\gamma},\ \boldsymbol{\theta}_0)] = \mathbf{0},$$

with $\nu_i = \mathbf{z}_{1i}^\mathsf{T}\boldsymbol{\gamma} + \mu_i = \mathbb{E}[r_i|\mathbf{z}_i]$, which is a conditional expectation projection. The sample estimating equation becomes

$$o_\mathbb{P}\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{n}\sum_i \mathbf{m}(\mathbf{w}_i,\ \hat{\nu}_i - \mathbf{z}_{1i}^\mathsf{T}\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\theta}}),$$

where both $\hat{\nu}_i$ and $\hat{\boldsymbol{\gamma}}$ are estimated by linear regression in a first step and are plugged into the second step estimating equation, from which then $\hat{\boldsymbol{\theta}}$ is obtained. We show in this section that introducing the additional parameter $\boldsymbol{\gamma}$ in the first step only affects the asymptotic variance of $\hat{\boldsymbol{\theta}}$, but not its bias properties. In particular, our theory on asymptotic bias with many covariates entering the first step remains unchanged with the first step now taking the partially linear form described above.

Under regularity conditions, $\hat{\boldsymbol{\gamma}}$ is $\sqrt{n}$-consistent for γ, and standard linearization arguments show that $\hat{\boldsymbol{\theta}}$ has the following representation:

$$\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big) = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i, \hat{\nu}_i - \mathbf{z}_{1i}^\mathsf{T}\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\gamma}}, \boldsymbol{\theta}_0) + o_\mathbb{P}(1).$$

Given that $\hat{\boldsymbol{\gamma}}$ is consistent, it is not hard to show the following:

$$\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big) = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\nu}_i - \mathbf{z}_{1i}^\mathsf{T}\boldsymbol{\gamma}, \boldsymbol{\gamma}, \boldsymbol{\theta}_0) + \boldsymbol{\Sigma}_0 \boldsymbol{\Xi}_0 \sqrt{n}\Big(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\Big) + o_\mathbb{P}(1),$$

with

$$\boldsymbol{\Xi}_0 = \mathbb{E}\Big[- \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\mathbf{z}_{1i}^\mathsf{T} + \frac{\partial}{\partial \boldsymbol{\gamma}^\mathsf{T}} \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\Big],$$

where $\dot{\mathbf{m}}$ continue to denote the first partial derivative of $\mathbf{m}$ with respect to μ, and $\ddot{\mathbf{m}}$ the second derivative. The next step is to further expand, which gives

$$\begin{aligned}
\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big) &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0) && \text{(E.23)} \\
&+ \boldsymbol{\Sigma}_0 \boldsymbol{\Xi}_0 \sqrt{n}\Big(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\Big) && \text{(E.24)} \\
&+ \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\Big(\hat{\nu}_i - \nu_i\Big) && \text{(E.25)} \\
&+ \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\Big(\hat{\nu}_i - \nu_i\Big)^2 + o_\mathbb{P}(1). && \text{(E.26)}
\end{aligned}$$

(E.23) is the influence function had both $\boldsymbol{\gamma}$ and μ_i been observed, (E.24) is the total variance contribution from estimating $\boldsymbol{\gamma}$, and (E.25) also gives a variance contribution because ν_i is estimated. Finally, both (E.25) and (E.26) will lead to asymptotic bias under our many covariates assumption.

We first consider (E.25) and (E.26). Since $\hat{\nu}_i$ is constructed as linear projection, the same technique developed in Section SA-3 can be applied. Let $\boldsymbol{\Pi} = \mathbf{Z}(\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-1}\mathbf{Z}^\mathsf{T}$ be the projection matrix constructed from the "long vector" $\mathbf{z}_i = [\mathbf{z}_{1i}^\mathsf{T}, \mathbf{z}_{2i}^\mathsf{T}]^\mathsf{T}$ with $\mathbf{Z}$ the $n \times (d_\gamma + k)$ matrix stacking $\mathbf{z}_i$, and π_{ij} be a generic element of $\boldsymbol{\Pi}$. Then, the same conditions are enough to justify:

$$\begin{aligned}
\text{(E.25)} &= \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)|\mathbf{z}_i]\pi_{ij} \right) \varepsilon_i + \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + o_\mathbb{P}(1), \\
\text{(E.26)} &= \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + o_\mathbb{P}(1),
\end{aligned}$$

with $\mathbf{b}_{1,i} = \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\varepsilon_i|\mathbf{z}_i]$ and $\mathbf{b}_{2,ij} = \frac{1}{2}\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\varepsilon_j^2|\mathbf{z}_i, \mathbf{z}_j]$.

The new term (E.24) can be written as

$$\text{(E.24)} = \boldsymbol{\Sigma}_0 \boldsymbol{\Xi}_0 \Big(\frac{1}{n}\mathbf{Z}_1^\mathsf{T}\mathbf{Q}_2\mathbf{Z}_1\Big)^{-1} \frac{1}{\sqrt{n}} \sum_i \Big(\sum_j \mathbf{z}_{1j} q_{2ij} \Big)\varepsilon_i + o_\mathbb{P}(1),$$

where $\mathbf{Z}_1$ is the $n \times d_\gamma$ matrix stacking $\mathbf{z}_{1i}$, $\mathbf{Q}_2$ is the $n \times n$ annihilator $\mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2^\mathsf{T}\mathbf{Z}_2)^{-1}\mathbf{Z}_2^\mathsf{T}$ to partial out $\mathbf{z}_{2i}$ with elements denoted by q_{2ij}, and the extra $o_\mathbb{P}(1)$ arises due to the approximation error η_i. Therefore, the extra term can also be shown to be asymptotically normal conditional on $\mathbf{Z}$.

Now we briefly mention regularity conditions that are used to justify the $\sqrt{n}$-consistency and conditional asymptotic normality of $\hat{\boldsymbol{\gamma}}$. The main reference is Cattaneo, Jansson and Newey (2018*b*), who established asymptotic normality of $\hat{\boldsymbol{\gamma}}$ in a much more general setting. We provide primitive conditions to verify their Assumptions 1–3.

We first need to modify the notation of Hölder continuity since now an additional nuisance parameter $\boldsymbol{\gamma}$ is allowed to enter the moment function directly. We re-define:

$$\mathcal{H}_i^{\alpha,\delta}(\mathbf{m}) = \sup_{(|\mu-\mu_i|+|\boldsymbol{\gamma}'-\boldsymbol{\gamma}|+|\boldsymbol{\theta}-\boldsymbol{\theta}_0|)^\alpha \leq \delta} \frac{|\mathbf{m}(\mathbf{w}_i,\mu,\boldsymbol{\gamma}',\boldsymbol{\theta}) - \mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\gamma},\boldsymbol{\theta}_0)|}{(|\mu-\mu_i|+|\boldsymbol{\gamma}'-\boldsymbol{\gamma}|+|\boldsymbol{\theta}-\boldsymbol{\theta}_0|)^\alpha}.$$

The same transformation is also applied to derivatives of the moment function.

Assumption (Partially Linear First Step)**.**

A.PL(1) The minimum eigenvalue of $\mathbb{V}[\mathbf{z}_{1i}|\mathbf{z}_{2i}]$ is bounded away from zero.

A.PL(2) $\mathbb{E}[|\mathbf{z}_{1i}|^4|\mathbf{z}_{2i}]$ is bounded.

A.PL(3) Both $\partial\mathbf{m}_i/\partial\boldsymbol{\gamma}$ and $\mathcal{H}_i^{\alpha,\delta}(\partial\mathbf{m}/\partial\boldsymbol{\gamma}) \in \mathsf{BM}_1$ have finite mean, for some α, $\delta > 0$. ⌟

The first requirement is intuitive: the high-dimensional vector $\mathbf{z}_{2i}$ is partialled out, there should be residual variation in $\mathbf{z}_{1i}$ so that $\boldsymbol{\gamma}$ is identified (consistently estimable). The second requirement imposes moment conditions.

In Cattaneo, Jansson and Newey (2018*b*), it is also assumed that $\mathbb{V}[\varepsilon_i|\mathbf{z}_{1i},\mathbf{z}_{2i}]$ is bounded away from zero. This is necessary to establish asymptotic normality of $\hat{\boldsymbol{\gamma}}$, since otherwise the asymptotic distribution could be degenerate. This condition, however, is not essential for our purpose: our target parameter is $\hat{\boldsymbol{\theta}}$, which has the expansion

$$\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{\boldsymbol{\mathcal{B}}}{\sqrt{n}}\Big) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_{\mathbb{P}}(1),$$

with

$$\begin{aligned}
\boldsymbol{\mathcal{B}} &= \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \left[\sum_i \mathbf{b}_{1,i}\pi_{ii} + \mathbf{b}_{2,ij}\pi_{ij}^2\right], \\
\bar{\boldsymbol{\Psi}}_1 &= \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \left[\sum_i \mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\gamma},\boldsymbol{\theta}_0)\right], \\
\bar{\boldsymbol{\Psi}}_2 &= \boldsymbol{\Sigma}_0\boldsymbol{\Xi}_0\Big(\mathbf{Z}_1^\mathsf{T}\mathbf{Q}_2\mathbf{Z}_1\Big)^{-1}\frac{1}{\sqrt{n}}\sum_i\Big(\sum_j \mathbf{z}_{1j}q_{2ij}\Big)\varepsilon_i + \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\sum_i\left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\gamma},\boldsymbol{\theta}_0)|\mathbf{z}_i]\pi_{ij}\right)\varepsilon_i,
\end{aligned}$$

and therefore the condition we need is that, with probability approaching one, the minimum eigenvalue of the "overall" variance-covariance matrix be bounded away from zero. See Theorem SA.6.

SA-4.3 Second Step: Additional Many Covariates

It is very difficult to extend Theorem SA.6 in full generality to a setting where both estimation steps are high-dimensional. Given that different asymptotic behaviors will emerge for the main estimator of interest $\hat{\boldsymbol{\theta}}$ in cases where the second step estimating equation includes high-dimensional covariates (and hence high-dimensional parameters that need to be estimated), it is natural to restrict the way the high-dimensional covariates enter the second step estimation procedure. In this section, we make a first attempt to generalize the problem so that the second step also has increasing dimension by imposing a particular restriction on the estimating equation for $\boldsymbol{\theta}$. Specifically, we consider a setting where the first step estimate $\hat{\mu}_i$ enters a high-dimensional semi-linear regression problem. We show that new biases arise as a result of both steps being high-dimensional, thereby showing that the main conclusions of our paper continue to hold in this case.

Let y_i be a response variable and assume that

$$\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i, \mu_i] = f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0) + \mathbf{z}_i^\mathsf{T}\boldsymbol{\gamma}_0,$$

where $\boldsymbol{\theta}_0$ is the parameter of interest and f is a known smooth function. We assume $\mathbf{x}_i$ has fixed dimension, and $\mathbf{z}_i$ has possibly increasing dimension but satisfies $k = O(\sqrt{n})$. This setting is a special case of our generic framework in that a non-linear least squares estimating equation is considered, but is also more general because a possibly high-dimensional vector of covariates is now allowed for in the second step, where the additional possibly high-dimensional parameter $\boldsymbol{\gamma}_0$ features. For simplicity, we assume the same high dimensional vector $\mathbf{z}_i$ is used in both the first step (to construct $\hat{\mu}_i$) and the second step (as additional controls).

Given this setting, and assuming a non-linear least-squares model is considered, we can map this problem into a generalization of our framework, as follows:

$$\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0) \\ \mathbf{z}_i \end{bmatrix} \Big(y_i - f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0) - \mathbf{z}_i^\mathsf{T}\boldsymbol{\gamma}_0 \Big),$$

with $\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)] = \mathbf{0}$, $\mathbf{w}_i = [y_i, \mathbf{x}_i^\mathsf{T}, \mathbf{z}_i^\mathsf{T}]^\mathsf{T}$, where $\boldsymbol{\theta}_0$ continues to be the parameter of interest and now the additional (possibly high-dimensional) parameter $\boldsymbol{\gamma}_0$ features in the second step estimating equation $\mathbf{m}(\cdot)$. Importantly, the second step estimating equation has increasing dimension and therefore is outside the scope of our paper.

Nevertheless, because we consider a semilinear least squares estimation problem, we can recast the second step by partialling out the increasing dimension covariates $\mathbf{z}_i$, and then exploit our main results along with natural extensions thereof. To describe this approach, we need additional notation. First, let $\mathbf{f} = \partial f/\partial \boldsymbol{\theta}$ be the derivative of f with respect to $\boldsymbol{\theta}$, and set $f_i = f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0)$ and $\mathbf{f}_i = \mathbf{f}(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0)$ to simplify notation. Second, let $q_{ij} = \mathbb{1}_{\{i=j\}} - \pi_{ij}$ denote the elements of the annihilation matrix $\mathbf{Q} = \mathbf{I} - \boldsymbol{\Pi}$. Finally, the "dot" notation is reserved for partial derivatives with respect to μ as done so far.

Regressing out the high-dimensional vector $\mathbf{z}_i$, the estimator of interest $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{\theta}} \quad : \quad \mathbf{0} = \sum_i \Big(\sum_j q_{ij} \mathbf{f}(\mathbf{x}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) \Big) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) \Big), \tag{E.27}$$

where $\hat{\mu}_i$ is constructed from projecting r_i on $\mathbf{z}_i$ in a first step as in our basic framework. Due to the presence of the high-dimensional vector in the second step, the asymptotic expansion becomes much more cumbersome. Details are relegated to Section SA-9.9, where we also discuss additional regularity conditions. To give the intuition of the result, note that a first order linearization with respect to $\hat{\mu}_i$ and consistency of the resulting Hessian matrix gives:

$$\begin{aligned}\sqrt{n}\Big(\mathbb{E}\mathbb{V}[\mathbf{f}_i|\mathbf{z}_i] + o_{\mathbb{P}}(1) \Big) \Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \Big) &= \frac{1}{\sqrt{n}} \sum_i \mathbf{f}(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) \\ &\quad - \frac{1}{\sqrt{n}} \sum_i \Big(\sum_j \pi_{ij} \mathbf{f}(\mathbf{x}_j, \hat{\mu}_j, \boldsymbol{\theta}_0) \Big) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) + o_{\mathbb{P}}(1).\end{aligned}$$

Comparing with the expansion in Section SA-3, an extra term emerges in the current setting due to the possibly high-dimensional covariates $\mathbf{z}_i$ entering (linearly) in the second step. Specifically, the first term after the equal sign can be analyzed using the results in Section SA-3 without modifications: simply set $\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = \mathbf{f}(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0)(y_i - f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0))$. The second term is the new term emerging from the linear projection out of the covariates $\mathbf{z}_i$ in the second step estimating equation: Section SA-9.9 discusses how this new term can be handled when $k = O(\sqrt{n})$.

In sum, under regularity conditions, we obtain the following stochastic expansion for the estimator $\hat{\boldsymbol{\theta}}$ defined in (E.27) when $k = O(\sqrt{n})$:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{\boldsymbol{\mathcal{B}}}{\sqrt{n}} \right) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_{\mathbb{P}}(1),$$

with $\boldsymbol{\Sigma}_0 = \mathbb{E}[\mathbb{V}[\mathbf{f}_i|\mathbf{z}_i]]^{-1}$, $u_i = y_i - \mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i, \mu_i]$, and

$$\begin{aligned}\boldsymbol{\mathcal{B}} &= \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \left[\sum_i \mathbf{b}_{1,i} \pi_{ii} + \sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2 + \sum_{i,j} \mathbf{b}_{3,ij} \pi_{ij}^3 + \sum_{i,j,\ell} \mathbf{b}_{4,ij\ell} \pi_{ij} \pi_{i\ell} \pi_{j\ell} \right] \\ \mathbf{b}_{1,i} &= \mathbb{E}[\dot{\mathbf{f}}_i u_i \varepsilon_i | \mathbf{z}_i] - \mathbb{C}\mathrm{ov}[\mathbf{f}_i, \dot{f}_i \varepsilon_i | \mathbf{z}_i] \\ \mathbf{b}_{2,ij} &= \mathbb{E}[\dot{f}_i|\mathbf{z}_i]\mathbb{E}[\mathbf{f}_j \varepsilon_j|\mathbf{z}_j] - \mathbb{E}[\dot{\mathbf{f}}_i|\mathbf{z}_i]\mathbb{E}[u_j \varepsilon_j|\mathbf{z}_j] - \frac{1}{2}\mathbb{C}\mathrm{ov}[\mathbf{f}_i, \ddot{f}_i|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_i, \mathbf{z}_j] - \mathbb{E}[\dot{\mathbf{f}}_i \dot{f}_i|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j] \\ \mathbf{b}_{3,ij} &= \frac{1}{2}\mathbb{E}[\ddot{f}_i|\mathbf{z}_i]\mathbb{E}[\mathbf{f}_j \varepsilon_j^2|\mathbf{z}_j] - \frac{1}{2}\mathbb{E}[\ddot{\mathbf{f}}_i|\mathbf{z}_i]\mathbb{E}[u_j \varepsilon_j^2|\mathbf{z}_j] + \mathbb{E}[\dot{f}_i \varepsilon_i|\mathbf{z}_i]\mathbb{E}[\dot{\mathbf{f}}_j \varepsilon_j|\mathbf{z}_j] \\ \mathbf{b}_{4,ij\ell} &= \mathbb{E}[\dot{f}_i|\mathbf{z}_i]\mathbb{E}[\dot{\mathbf{f}}_\ell|\mathbf{z}_\ell]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j],\end{aligned}$$

and

$$\bar{\boldsymbol{\Psi}}_1 = \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_i (\mathbf{f}_i - \mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]) u_i, \qquad \bar{\boldsymbol{\Psi}}_2 = -\frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_i \mathbb{C}\mathrm{ov}[\mathbf{f}_i, \dot{f}_i|\mathbf{z}_i] \varepsilon_i.$$

This result is obtained under Assumptions A.2 and $\mathbb{E}[|\zeta_i|^2] = o(1)$, but the later misspecification/smoothing bias restriction is imposed for simplicity and could be dropped. The resulting stochastic expansion shows that including high-dimensional covariates in the second step estimation, when entering linearly and using non-linear least squares for estimation, leads to the presence of a many covariates bias of the same order as reported in our main Theorem SA.6. The specific form of the bias changes because of the interaction between the first and second step estimation, as now both include high-dimensional covariates. The non-linearity introduced by the first step estimate entering the second step estimating equation plays a crucial role in this result. Because least squares estimators are *not* linear in covariates, which means that the many covariates bias emerges even when μ_i enters the second step multiplicatively (i.e. f is linear in μ_i). On the other hand, it is known that one-step high-dimensional linear least squares estimators will not lead to a many covariates bias as shown in Cattaneo, Jansson and Newey (2018*a*,*b*), which is due to the intrinsic linearity of that setting. We now discuss a few specialized examples to illustrate some implications of this extension.

The effect of high dimensional second step

Although we only consider a special case of high dimensional second step, one can already see some of its implications. To compare, consider what happens if $\boldsymbol{\gamma}_0 = \mathbf{0}$ and the long vector $\mathbf{z}_i$ is excluded from the second step, i.e.,

$$\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i, \mu_i] = f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0),$$

and denote by $\hat{\boldsymbol{\theta}}$ the estimator obtained using non-linear least squares. Then, our main Theorem SA.6 applies directly, and gives

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{\boldsymbol{\mathcal{B}}}{\sqrt{n}}\right) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_{\mathbb{P}}(1),$$

with $\boldsymbol{\Sigma}_0 = \mathbb{E}[\mathbf{f}_i\mathbf{f}_i^{\mathsf{T}}]^{-1}$, and

$$\begin{aligned}
\boldsymbol{\mathcal{B}} &= \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\left[\sum_i \mathbf{b}_{1,i}\pi_{ii} + \sum_{i,j}\mathbf{b}_{2,ij}\pi_{ij}^2\right] \\
\mathbf{b}_{1,i} &= \mathbb{E}[\dot{\mathbf{f}}_i u_i\varepsilon_i|\mathbf{z}_i] - \mathbb{E}[\mathbf{f}_i\dot{f}_i\varepsilon_i|\mathbf{z}_i], \qquad \mathbf{b}_{2,ij} = -\frac{1}{2}\mathbb{E}[\mathbf{f}_i\ddot{f}_i|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j] - \mathbb{E}[\dot{\mathbf{f}}_i\dot{f}_i|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j] \\
\bar{\boldsymbol{\Psi}}_1 &= \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\sum_i \mathbf{f}_i u_i, \qquad \bar{\boldsymbol{\Psi}}_2 = -\frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\sum_i \mathbb{E}[\mathbf{f}_i\dot{f}_i|\mathbf{z}_i]\varepsilon_i.
\end{aligned}$$

Hence including the high dimensional control variables $\mathbf{z}_i$ in the second step has two effects. First, additional bias terms arise. Second, some conditional expectations in the expansion become conditional covariances, since $\mathbf{z}_i$ has to be partialed out from f.

Special Case 1: high dimensional regression with generated regressor

Assume now that the second step becomes a regression problem:

$$\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i, \mu_i] = g(\mathbf{x}_i, \mu_i) \cdot \theta_0 + \mathbf{z}_i^\mathsf{T} \boldsymbol{\gamma}_0,$$

which means $f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}) = g(\mathbf{x}_i, \mu_i) \cdot \theta$. Then we can set $\mathbf{f} = g$ and $f = \theta_0 g$ in earlier results, which implies the following bias and variance formula

$$\sqrt{n}\left(\hat{\theta} - \theta_0 - \frac{\mathcal{B}}{\sqrt{n}}\right) = \bar{\Psi}_1 + \bar{\Psi}_2 + o_{\mathbb{P}}(1),$$

with $\Sigma_0 = \mathbb{E}[\mathbb{V}[g_i|\mathbf{z}_i]]^{-1}$

$$\begin{aligned}
\mathcal{B} &= \frac{1}{\sqrt{n}}\Sigma_0 \left[\sum_i b_{1,i}\pi_{ii} + \sum_{i,j} b_{2,ij}\pi_{ij}^2 + \sum_{i,j} b_{3,ij}\pi_{ij}^3 + \sum_{i,j,\ell} b_{4,ij\ell}\pi_{ij}\pi_{i\ell}\pi_{j\ell}\right] \\
b_{1,i} &= \mathbb{E}[\dot{g}_i u_i \varepsilon_i|\mathbf{z}_i] - \theta_0 \mathbb{C}\mathrm{ov}[g_i, \dot{g}_i\varepsilon_i|\mathbf{z}_i] \\
b_{2,ij} &= \theta_0 \mathbb{E}[\dot{g}_i|\mathbf{z}_i]\mathbb{E}[g_j\varepsilon_j|\mathbf{z}_j] - \mathbb{E}[\dot{g}_i|\mathbf{z}_i]\mathbb{E}[u_j\varepsilon_j|\mathbf{z}_j] - \frac{\theta_0}{2}\mathbb{C}\mathrm{ov}[g_i, \ddot{g}_i|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_i, \mathbf{z}_j] - \theta_0\mathbb{E}[\dot{g}_i^2|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j] \\
b_{3,ij} &= \frac{\theta_0}{2}\mathbb{E}[\ddot{g}_i|\mathbf{z}_i]\mathbb{E}[g_j\varepsilon_j^2|\mathbf{z}_j] - \frac{1}{2}\mathbb{E}[\ddot{g}_i|\mathbf{z}_i]\mathbb{E}[u_j\varepsilon_j^2|\mathbf{z}_j] + \theta_0\mathbb{E}[\dot{g}_i\varepsilon_i|\mathbf{z}_i]\mathbb{E}[\dot{g}_j\varepsilon_j|\mathbf{z}_j] \\
b_{4,ij\ell} &= \theta_0\mathbb{E}[\dot{g}_i|\mathbf{z}_i]\mathbb{E}[\dot{g}_\ell|\mathbf{z}_\ell]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j],
\end{aligned}$$

and

$$\bar{\Psi}_1 = \frac{1}{\sqrt{n}}\Sigma_0 \sum_i (g_i - \mathbb{E}[g_i|\mathbf{z}_i])u_i, \qquad \bar{\Psi}_2 = -\frac{1}{\sqrt{n}}\Sigma_0 \sum_i \theta_0 \mathbb{C}\mathrm{ov}[g_i, \dot{g}_i|\mathbf{z}_i]\varepsilon_i.$$

Both the variance and the bias remain essentially the same.

Special Case 2: multiplicative μ_i

An even more special case is the following

$$\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i, \mu_i] = (x_i\mu_i) \cdot \theta_0 + \mathbf{z}_i^\mathsf{T} \boldsymbol{\gamma}_0,$$

so that $f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}) = x_i\mu_i \cdot \theta$. Now it seems the asymptotic bias should vanish since the generated regressor $\hat{\mu}_i$ enters the second step multiplicatively. Unfortunately, linear regression is not linear in the regressors, and the many covariates bias remains. (Although some of the terms in the expansion do disappear.) Corresponding results can be obtained with $g_i = x_i\mu_i$, $\dot{g}_i = x_i$ and $\ddot{g} = 0$, following the notation from the previous special case. Hence,

$$\sqrt{n}\left(\hat{\theta} - \theta_0 - \frac{\mathcal{B}}{\sqrt{n}}\right) = \bar{\Psi}_1 + \bar{\Psi}_2 + o_{\mathbb{P}}(1),$$

with $\Sigma_0 = \mathbb{E}[\mu_i^2 \mathbb{V}[x_i|\mathbf{z}_i]]^{-1}$

$$\begin{aligned}
\mathcal{B} &= \frac{1}{\sqrt{n}}\Sigma_0 \left[\sum_i b_{1,i}\pi_{ii} + \sum_{i,j} b_{2,ij}\pi_{ij}^2 + \sum_{i,j} b_{3,ij}\pi_{ij}^3 + \sum_{i,j,\ell} b_{4,ij\ell}\pi_{ij}\pi_{i\ell}\pi_{j\ell} \right] \\
b_{1,i} &= \mathbb{E}[x_i u_i \varepsilon_i|\mathbf{z}_i] - \theta_0 \mu_i \mathbb{C}\mathrm{ov}[x_i, x_i\varepsilon_i|\mathbf{z}_i] \\
b_{2,ij} &= \theta_0 \mu_j \mathbb{E}[x_i|\mathbf{z}_i]\mathbb{E}[x_j\varepsilon_j|\mathbf{z}_j] - \mathbb{E}[x_i|\mathbf{z}_i]\mathbb{E}[u_j\varepsilon_j|\mathbf{z}_j] - \theta_0 \mathbb{E}[x_i^2|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j] \\
b_{3,ij} &= \theta_0 \mathbb{E}[x_i\varepsilon_i|\mathbf{z}_i]\mathbb{E}[x_j\varepsilon_j|\mathbf{z}_j] \\
b_{4,ij\ell} &= \theta_0 \mathbb{E}[x_i|\mathbf{z}_i]\mathbb{E}[x_\ell|\mathbf{z}_\ell]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j],
\end{aligned}$$

and

$$\bar{\Psi}_1 = \frac{1}{\sqrt{n}}\Sigma_0 \sum_i (x_i - \mathbb{E}[x_i|\mathbf{z}_i])\mu_i u_i, \qquad \bar{\Psi}_2 = -\frac{1}{\sqrt{n}}\Sigma_0 \sum_i \theta_0 \mathbb{V}[x_i|\mathbf{z}_i]\mu_i \varepsilon_i.$$

The above result indeed confirms that the many covariates bias remains to be present even in a simple problem where the estimated $\hat{\mu}_i$ is used as a regressor.

SA-5 Examples

Our results cover a wide range of applications in econometrics and statistics. In this section, we show that overfitting the first step estimate can generate a first order bias contribution for many estimators commonly used in practice. We give exact bias and variance formulas for several methods in treatment effect, policy evaluation and other applied microeconomic settings (see Abadie and Cattaneo, 2018, for review and further references). For many of these examples, it is possible to give intuition for the source of the many covariates bias by combining the general bias formula with the specific (identification) assumptions imposed.

Notation. To avoid notation conflicts and confusion, we use uppercase letters to denote random variables in this section, such as X_i, W_i, etc. Random vectors will be denoted by bold upper case letters, such as $\mathbf{X}_i$, $\mathbf{W}_i$, etc. This should be distinguished from matrices, where the latter are not indexed by i, such as $\mathbf{A}$, $\mathbf{Z}$, etc.

SA-5.1 Inverse Probability Weighting

We consider estimation via IPW in a general missing data problem. Our results also apply immediately to treatment effect, data combination, and measurement error settings, when a conditional independence assumption is imposed. Assume the parameter of interest is $\mathbb{E}[\mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$, where $\mathbf{Y}_i(1)$ is subject to a missing data problem and $\mathbf{X}_i$ are covariates of fixed dimension. Let $T_i = \mathbb{1}[\mathbf{Y}_i(1) \text{ is observed}]$, then $\mathbf{Y}_i = T_i\mathbf{Y}_i(1)$ is the observed vector. Under the assumptions below,

the parameter $\boldsymbol{\theta}_0$ is identifiable by the following estimating equation

$$\mathbb{E}\left[\frac{T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i}\right] = \mathbf{0},$$

and $P_i = \mathbb{E}[T_i|\mathbf{Z}_i]$ is the propensity score. We assume $\mathbf{X}_i$ is a subvector of $\mathbf{Z}_i$.

Assumption (IPW).

A.IPW(1) $\boldsymbol{\theta}_0$ is the unique root of $\mathbb{E}[\mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta})]$.

A.IPW(2) There exists C, such that $0 < C \leq P_i = \mathbb{E}[T_i|\mathbf{Z}_i]$.

A.IPW(3) Conditional on $\mathbf{Z}_i$, T_i is independent of $\mathbf{Y}_i(1)$. ⌟

For simplicity, we also assume that the dimension of $\mathbf{h}$ is the same as that of $\boldsymbol{\theta}$, hence the parameter is exactly identified, which implies we do not need to use an extra weighting matrix. The estimator is then defined by the two step procedure:

$$\frac{1}{\sqrt{n}} \sum_i \frac{T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0)}{\hat{P}_i} = \mathbf{0},$$

and $\hat{P}_i$ is the linear projection of T_i on $\mathbf{Z}_i$.

This example fits our framework as follows:

$$\mathbf{w}_i = [\mathbf{Y}_i^\mathsf{T}, \mathbf{X}_i^\mathsf{T}, T_i]^\mathsf{T},\ r_i = T_i,\ \mu_i = P_i,\ \mathbf{z}_i = \mathbf{Z}_i,\ \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}) = T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta})/P_i.$$

Applying Theorem SA.5, we have the following:

Proposition SA.7 (IPW).
Suppose the assumptions of Theorem SA.5 and IPW hold. Then, the IPW estimator $\hat{\boldsymbol{\theta}}$ is consistent, and admits the following representation:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{\boldsymbol{\mathcal{B}}}{\sqrt{n}}\right) = \bar{\boldsymbol{\Psi}} + o_{\mathbb{P}}(1),$$

where $\mathbf{g}_i = \mathbb{E}[\mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0)|\mathbf{Z}_i]$, and

$$\begin{aligned}
\boldsymbol{\mathcal{B}} &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \left[-\sum_i \mathbf{g}_i \frac{1 - P_i}{P_i} \pi_{ii} + \sum_{i,j} \mathbf{g}_i \frac{\mathbb{E}[T_i \varepsilon_j^2 | \mathbf{Z}_i, \mathbf{Z}_j]}{P_i^3} \pi_{ij}^2 \right] \\
\bar{\boldsymbol{\Psi}} &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \left[\frac{T_i \mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i} - \left(\sum_j \frac{\mathbf{g}_j}{P_j} \pi_{ij} \right) \cdot \varepsilon_i \right] \\
\boldsymbol{\Sigma}_0 &= \left(-\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0) \right] \right)^{-1}.
\end{aligned}$$

If, in addition, $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(1)$, then

$$\bar{\boldsymbol{\Psi}} = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \left[\frac{T_i \mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i} - \frac{\mathbf{g}_i}{P_i} \cdot \varepsilon_i \right] + o_{\mathbb{P}}(1).$$

The bias will be zero in this example, if either: (i) $P_i = 1$, which implies there is no missing values in the sample, or (ii) $\mathbf{g}_i = \boldsymbol{\theta}_0$, so that $\mathbf{Z}_i$ does not enter the outcome equation. Neither of these conditions are likely to hold in practice, hence bias will be a concern if IPW methods with overfitted propensity score are used. In addition, it follows that two assumptions are needed to achieve semiparametric efficiency. First, $k = o(\sqrt{n})$ so that the specification of the propensity score has to be relatively parsimonious. Second, the covariates $\mathbf{Z}_i$ must have approximation power for $\dot{\mathbf{m}}_i = -\mathbf{g}_i/P_i$.

Finally, we provide the following corollary, which specializes the previous conclusion to the case where only an outcome variable is missing and the goal is to estimate its mean $\theta_0 = \mathbb{E}[Y_i(1)]$.

Corollary SA.8 (IPW: Estimating Mean of $Y_i(1)$)**.**
Let $\mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0) = Y_i - \theta_0$. Suppose the assumptions of Theorem SA.5 and IPW hold. Then, the estimator $\hat{\theta}$ is consistent, and admits the following representation:

$$\sqrt{n}\left(\hat{\theta} - \theta_0 - \frac{\mathcal{B}}{\sqrt{n}}\right) = \bar{\Psi} + o_{\mathbb{P}}(1),$$

where $g_i = \mathbb{E}[Y_i(1)|\mathbf{z}_i] - \theta_0$, and

$$\mathcal{B} = \frac{1}{\sqrt{n}} \left[-\sum_i \frac{(1-P_i)g_i}{P_i} \pi_{ii} + \sum_{i,j} g_i \frac{\mathbb{E}[T_i \varepsilon_j^2 | \mathbf{Z}_i, \mathbf{Z}_j]}{P_i^3} \pi_{ij}^2 \right]$$

$$\bar{\Psi} = \frac{1}{\sqrt{n}} \sum_i \left[\frac{T_i(Y_i(1) - \theta_0)}{P_i} - \left(\sum_j \frac{g_j}{P_j} \pi_{ij} \right) \cdot (T_i - P_i) \right].$$

If, in addition, $\mathbb{E}[|\zeta_i|^2] = o(1)$, then

$$\bar{\Psi} = \frac{1}{\sqrt{n}} \sum_i \left[\frac{T_i(Y_i(1) - \theta_0)}{P_i} - \frac{g_i}{P_i}(T_i - P_i) \right] + o_{\mathbb{P}}(1).$$

⌟

SA-5.2 Semiparametric Difference-in-Differences

Suppose for each individual i two observations are available in two time periods, t_1 and t_2, which we denote by $Y_i(t_1)$ and $Y_i(t_2)$, respectively. Assume at time t_2 some individuals receive a treatment, with $T_i = 1$ the treatment indicator. In a potential outcomes framework, the second period outcome is $Y_i(t_2) = Y_i(1, t_2)T_i + Y_i(0, t_2)(1 - T_i)$, where $(Y_i(1, t_2), Y_i(0, t_2))$ are the potential outcomes in the second period tracking whether unit i received treatment or not, respectively. The parameter of

interest is the average treatment on the treated in the second period:

$$\theta_0 = \mathbb{E}[Y_i(1,t_2) - Y_i(0,t_2)|T_i = 1].$$

A classical identification assumption is the so-called "parallel trends" assumption. Abadie (2005) relaxes this assumption to "parallel trends conditional on covariates":

Assumption (DiD)**.**

A.DiD(1) $\mathbb{E}[Y_i(0,t_2) - Y_i(t_1)|T_i = 1, \mathbf{X}_i] = \mathbb{E}[Y_i(0,t_2) - Y_i(t_1)|T_i = 0, \mathbf{X}_i]$.

A.DiD(2) There exists $0 < C < 1$, such that $C \leq \mathbb{P}[T_i = 1|\mathbf{X}_i] \leq 1 - C$. ⌟

Assumptions A.DiD(1) and A.DiD(2), and regularity conditions such as bounded moments, imply

$$\theta_0 = \mathbb{E}\left[\frac{T_i - P_i}{\mathbb{P}[T_i = 1] \cdot (1 - P_i)}\Big(Y_i(t_2) - Y_i(t_1)\Big)\right],$$

where $P_i = \mathbb{P}[T_i = 1|\mathbf{X}_i]$ is the propensity score, and therefore θ_0 is identifiable from the marginal distribution of the observed quantities.

To fit this example into our framework, define:

$$\mathbf{w}_i = [T_i, Y_i(\cdot)]^\mathsf{T},\ r_i = T_i,\ \mu_i = P_i,\ \mathbf{z}_i = \text{series expansion of } \mathbf{X}_i$$
$$\mathbf{m}(\mathbf{w}_i, \mu_i, \theta) = \frac{T_i - P_i}{1 - P_i}\Big(Y_i(t_2) - Y_i(t_1)\Big) - T_i\theta.$$

We have the following result, which is Theorem SA.5 specialized to this model.

Proposition SA.9 (DiD)**.**
Suppose the assumptions of Theorem SA.5 and DiD hold. Then, the semiparametric difference-in-differences estimator $\hat{\theta}$ is consistent, and admits the following representation:

$$\sqrt{n}\left(\hat{\theta} - \theta_0 - \frac{\mathcal{B}}{\sqrt{n}}\right) = \bar{\Psi} + o_{\mathbb{P}}(1),$$

where $g_i = \mathbb{E}\left[Y_i(0,t_2) - Y_i(t_1)\middle| T_i = 1, \mathbf{X}_i\right] = \mathbb{E}\left[Y_i(0,t_2) - Y_i(t_1)\middle| T_i = 0, \mathbf{X}_i\right]$, and

$$\mathcal{B} = \frac{1}{\sqrt{n}\mathbb{P}[T_i = 1]}\left[\sum_i \frac{P_i}{1 - P_i} g_i \pi_{ii} - \sum_{i,j} \frac{\mathbb{E}[(T_j - P_j)^2|T_i = 0, \mathbf{X}_i, \mathbf{X}_j]}{(1 - P_i)^2} g_i \pi_{ij}^2\right]$$
$$\bar{\Psi} = \frac{1}{\sqrt{n}\mathbb{P}[T_i = 1]}\sum_i \left[\frac{T_i - P_i}{1 - P_i}\Big(Y_i(t_2) - Y_i(t_1)\Big) - T_i\theta_0 - \left(\sum_j \frac{1}{1 - P_j} g_j \pi_{ij}\right)\varepsilon_i\right].$$

If, in addition, $\mathbb{E}[|\zeta_i|^2] = o(1)$, then

$$\bar{\Psi} = \frac{1}{\sqrt{n}\mathbb{P}[T_i = 1]}\sum_i \left[\frac{T_i - P_i}{1 - P_i}\Big(Y_i(t_2) - Y_i(t_1)\Big) - T_i\theta_0 - \frac{1}{1 - P_i} g_i \varepsilon_i\right] + o_{\mathbb{P}}(1).$$

SA-5.3 Local Average Response Function

As a third example, we consider the semiparametric IV estimator of Abadie (2003). Let $D_i \in \{0,1\}$ be a (binary) instrumental variable (e.g., treatment assignment), and $T_i \in \{0,1\}$ be the (observed) treatment status indicator $T_i = D_i T_i(1) + (1-D_i)T_i(0)$, where $T_i(0)$ and $T_i(1)$ denote the two potential treatment statuses under control and treatment assignment, respectively. Let Y_i be the observed outcome: $Y_i = T_i D_i Y_i(1,1) + T_i(1-D_i)Y_i(1,0) + (1-T_i)D_i Y_i(0,1) + (1-T_i)(1-D_i)Y_i(0,0)$, where $Y_i(t,d)$ are four potential outcomes with their first and second arguments corresponding to treatment status ($t \in \{0,1\}$) and the value of the instrument ($d \in \{0,1\}$), respectively.

For identification, we rely on the following assumption, where $\mathbf{Z}_i$ are additional covariates.

Assumption (LARF)**.**

A.LARF(1) $\mathbb{P}[Y_i(0,0) = Y_i(0,1)|\mathbf{Z}_i] = 1$ and $\mathbb{P}[Y_i(1,0) = Y_i(1,1)|\mathbf{Z}_i] = 1$.

A.LARF(2) Conditional on $\mathbf{Z}_i$, $(T_i(0), T_i(1), Y_i(0), Y_i(1))$ are independent of D_i.

A.LARF(3) $C < \mathbb{P}[D_i = 1|\mathbf{Z}_i] < 1 - C$ for $0 < C < 1$, and $\mathbb{P}[T_i(1) = 1|\mathbf{Z}_i] > \mathbb{P}[T_i(0) = 1|\mathbf{Z}_i]$.

A.LARF(4) $\mathbb{P}[T_i(1) \geq T_i(0)|\mathbf{Z}_i] = 1$. ⌟

Assumption A.LARF(1) states that the instrumental variable D_i does not affect the outcome directly, hence it makes sense to use notation $Y_i(0)$, $Y_i(1)$ and $Y_i = T_i Y_i(1) + (1-T_i)Y_i(0)$ almost surely conditional on $\mathbf{Z}_i$. The second assumption, A.LARF(2) simply states the exogeneity of the instrument, which is standard in the literature. Assumption A.LARF(3) requires, after conditioning on $\mathbf{Z}_i$, there is variation in the instrument, which in turn indices variation in the treatment status. Finally, A.LARF(4) is typically referred as the monotonicity assumption, and it rules out defiers.

Letting $g(Y_i, T_i, \mathbf{Z}_i)$ be a known function with finite moments, Abadie (2003) showed that:

$$\mathbb{E}\Big[g(Y_i,T_i,\mathbf{Z}_i)\Big|T_i(1) > T_i(0)\Big] = \frac{\mathbb{E}[\kappa_i \cdot g(Y_i,T_i,\mathbf{Z}_i)]}{\mathbb{P}[T_i(1) > T_i(0)]}, \qquad \kappa_i = 1 - \frac{T_i(1-D_i)}{1-P_i} - \frac{(1-T_i)D_i}{P_i},$$

where $P_i = \mathbb{E}[D_i|\mathbf{Z}_i] = \mathbb{P}[D_i = 1|\mathbf{Z}_i]$. While compliers are not identifiable, this result allows identification of any statistical feature depending only on the joint distribution of $(Y_i, T_i, \mathbf{Z}_i)$ for compliers. Furthermore, this identification result allows for modelling the outcome variable Y_i with prespecified functional form (e.g., including pre-intervention covariates).

Suppose the parameter of interest is the conditional expectation function $\mathbb{E}[Y_i|T_i, \mathbf{X}_i, T_i(1) > T_i(0)] = e(\mathbf{X}_i, T_i, \boldsymbol{\theta})$, where $\mathbf{X}_i$ is a sub-vector of $\mathbf{Z}_i$ with fixed dimension d_x, and $e(\cdot, \boldsymbol{\theta})$ is known up to the finite dimensional parameter $\boldsymbol{\theta}$. Then, Abadie (2003)'s identification result implies

$$\mathbb{E}\Big[\kappa_i \cdot \frac{\partial e}{\partial \boldsymbol{\theta}}(\mathbf{X}_i, T_i, \boldsymbol{\theta})\Big(Y_i - e(\mathbf{X}_i, T_i, \boldsymbol{\theta})\Big)\Big] = \mathbf{0} \qquad \Leftrightarrow \qquad \boldsymbol{\theta} = \boldsymbol{\theta}_0,$$

which will be the (population) estimating equation for a non-linear least squares approach.

Replacing population quantities by their estimators, we fit this problem into our framework:

$$\mathbf{w}_i = [Y_i, T_i, D_i, \mathbf{X}_i^\mathsf{T}]^\mathsf{T},\ r_i = D_i,\ \mu_i = P_i,\ \mathbf{z}_i = \mathbf{Z}_i$$
$$\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}) = \kappa_i \cdot \frac{\partial e}{\partial \boldsymbol{\theta}}(\mathbf{X}_i, T_i, \boldsymbol{\theta})\Big(Y_i - e(\mathbf{X}_i, T_i, \boldsymbol{\theta})\Big).$$

Essentially, the above is a weighted nonlinear least squares problem where the weights have to be estimated. We have the following result, which is Theorem SA.5 specialized to the current context.

Proposition SA.10 (LARF)**.**
Supppose the assumptions of Theorem SA.5 and LARF hold. Then, $\hat{\boldsymbol{\theta}}$ is consistent, and admits the following representation:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{\boldsymbol{\mathcal{B}}}{\sqrt{n}}\right) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_\mathbb{P}(1),$$

where

$$\boldsymbol{\mathcal{B}} = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left[\sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2\right],$$
$$\bar{\boldsymbol{\Psi}}_1 = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0), \qquad \bar{\boldsymbol{\Psi}}_2 = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)|\mathbf{Z}_i\right] \cdot \pi_{ij}\right) \varepsilon_i,$$

and

$$\mathbf{b}_{1,i} = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0) \cdot \Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \cdot \Big(\frac{T_i D_i}{1 - P_i} + \frac{(1 - T_i)(1 - D_i)}{P_i}\Big) \middle| \mathbf{Z}_i, T_i(0) = T_i(1)\right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i],$$
$$\mathbf{b}_{2,ij} = \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0) \cdot \Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \cdot \Big(\frac{(1 - T_i)D_i}{P_i^3} + \frac{T_i(1 - D_i)}{(1 - P_i)^3}\Big)\varepsilon_j^2 \middle| \mathbf{Z}_i, \mathbf{Z}_j, T_i(0) = T_i(1)\right],$$
$$\cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i]$$
$$\mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)|\mathbf{Z}_j\right],$$
$$= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_j(\boldsymbol{\theta}_0) \cdot \Big(Y_j - e_j(\boldsymbol{\theta}_0)\Big)\left(\frac{1 - T_j}{P_j} - \frac{T_j}{1 - P_j}\right) \middle| \mathbf{Z}_j, T_j(0) = T_j(1)\right] \cdot \mathbb{P}\left[T_j(0) = T_j(1)|\mathbf{Z}_j\right]$$
$$e_i(\boldsymbol{\theta}) = e(\mathbf{X}_i, T_i, \boldsymbol{\theta}),$$
$$\boldsymbol{\Sigma}_0 = \left(-\mathbb{E}\left[\kappa_i \cdot \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) - \frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}^\mathsf{T}} e_i(\boldsymbol{\theta}_0)\right]\right]\right)^{-1}.$$

⌟

SA-5.4 Marginal Treatment Effect

This is one of the examples discussed in the main paper, which can be interpreted as a generalization of the local average response function when the instrument is not binary. First, we recall the basic

setup; see Heckman and Vytlacil (2005) and references therein for more details.

Employing again the potential outcomes framework, let Y_i be the outcome variable with $Y_i = T_iY_i(1)+(1-T_i)Y_i(0)$, where T_i is the treatment indicator. Given covariates $\mathbf{X}_i \in \mathbb{R}^{d_x}$, we decompose the potential outcomes into the conditional expectations and errors: $Y_i(0) = g_0(\mathbf{X}_i) + U_{0i}$ and $Y_i(1) = g_1(\mathbf{X}_i) + U_{1i}$. The (treatment) selection rule is $T_i = \mathbb{1}[P_i \geq V_i]$, where P_i is the propensity score (function of $\mathbf{Z}_i$) and V_i is uniformly distributed in $[0,1]$ conditional on $\mathbf{X}_i$.

The parameter of interest is the marginal treatment effect (MTE): $\tau(a|\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0)|V_i = a, \mathbf{X}_i = \mathbf{x}]$. This parameter is identifiable under the following assumption.

Assumption (MTE).

A.MTE(1) $\mathbf{X}_i \subset \mathbf{Z}_i$; and conditional on $\mathbf{X}_i$, P_i (and $\mathbf{Z}_i$) are nondegenerate and independent of (U_{1i}, U_{0i}, V_i).

A.MTE(2) $0 < \mathbb{P}[T_i = 1|\mathbf{X}_i] < 1$. ⌟

This assumption implies that, under regularity conditions,

$$\tau(a|\mathbf{x}) = \frac{\partial}{\partial p}\mathbb{E}[Y_i|P_i = p, \mathbf{X}_i = \mathbf{x}]\bigg|_{p=a}.$$

We further assume that $\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}] = e(\mathbf{x}, a, \boldsymbol{\theta}_0)$, where e is known function up to the unknown parameter $\boldsymbol{\theta}_0$, and therefore $\tau(a|\mathbf{x}) = \frac{\partial}{\partial p}e(\mathbf{x}, p, \boldsymbol{\theta}_0)|_{p=a}$. Now the problem can be framed into a general Z-estimation setup employing non-linear least squares:

$$\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}}e(\mathbf{X}_i, P_i, \boldsymbol{\theta})\Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta})\Big)\right] = \mathbf{0} \qquad \Leftrightarrow \qquad \boldsymbol{\theta} = \boldsymbol{\theta}_0,$$

Usually the parameter of interest is not $\boldsymbol{\theta}_0$, but rather the MTE curve or a weighted average thereof. As discussed in the main paper, we can rely on the delta method to apply our results to the latter estimands. Finally, because the propensity score P_i is unknown, it has to be estimated/approximated in a first step.

This problem also fits naturally into our general framework:

$$\mathbf{w}_i = [Y_i, \mathbf{X}_i^\mathsf{T}]^\mathsf{T},\ r_i = T_i,\ \mu_i = P_i,\ \mathbf{z}_i = \mathbf{Z}_i$$
$$\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}e(\mathbf{X}_i, P_i, \boldsymbol{\theta})\Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta})\Big).$$

Applying our Theorem SA.5, we obtain the following result. To save the notation, we set $e_i(\boldsymbol{\theta}) = e(\mathbf{X}_i, P_i, \boldsymbol{\theta})$, $\dot{e}_i(\boldsymbol{\theta}) = \frac{\partial}{\partial p}e(\mathbf{X}_i, p, \boldsymbol{\theta})|_{p=P_i}$, and $\ddot{e}_i(\boldsymbol{\theta}) = \frac{\partial^2}{\partial p^2}e(\mathbf{X}_i, p, \boldsymbol{\theta})|_{p=P_i}$.

Proposition SA.11 (MTE).
Suppose the assumptions of Theorem SA.5 and MTE hold. Then, $\hat{\boldsymbol{\theta}}$ is consistent, and admits the

following representation:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0-\frac{1}{\sqrt{n}}\boldsymbol{\mathcal{B}}\right)=\bar{\boldsymbol{\Psi}}_1+\bar{\boldsymbol{\Psi}}_2+o_{\mathbb{P}}(1),$$

where

$$\boldsymbol{\mathcal{B}}=\boldsymbol{\Sigma}_0\frac{1}{\sqrt{n}}\left[\sum_i \mathbf{b}_{1,i}\cdot\pi_{ii}+\sum_{i,j}\mathbf{b}_{2,ij}\cdot\pi_{ij}^2\right]$$

$$\bar{\boldsymbol{\Psi}}_1=\boldsymbol{\Sigma}_0\frac{1}{\sqrt{n}}\sum_i\frac{\partial}{\partial\boldsymbol{\theta}}e_i(\boldsymbol{\theta}_0)\Big(Y_i-e_i(\boldsymbol{\theta}_0)\Big),\qquad \bar{\boldsymbol{\Psi}}_2=\boldsymbol{\Sigma}_0\frac{-1}{\sqrt{n}}\sum_i\left(\sum_j\frac{\partial}{\partial\boldsymbol{\theta}}e_j(\boldsymbol{\theta}_0)\cdot\dot{e}_j(\boldsymbol{\theta}_0)\cdot\pi_{ij}\right)\varepsilon_i,$$

and

$$\begin{aligned}\mathbf{b}_{1,i}&=\frac{\partial}{\partial\boldsymbol{\theta}}\dot{e}_i(\boldsymbol{\theta}_0)\Big[(1-P_i)\cdot\mathbb{E}[T_iY_i(1)|\mathbf{Z}_i]-P_i\cdot\mathbb{E}[(1-T_i)Y_i(0)|\mathbf{Z}_i]\Big]\\ \mathbf{b}_{2,ij}&=-\frac{1}{2}\Big(2\frac{\partial}{\partial\boldsymbol{\theta}}\dot{e}_i(\boldsymbol{\theta}_0)\cdot\dot{e}_i(\boldsymbol{\theta}_0)+\frac{\partial}{\partial\boldsymbol{\theta}}e_i(\boldsymbol{\theta}_0)\cdot\ddot{e}_i(\boldsymbol{\theta}_0)\Big)P_j(1-P_j)\\ \boldsymbol{\Sigma}_0&=\left(\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}e_i(\boldsymbol{\theta}_0)\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}e_i(\boldsymbol{\theta}_0)\right]\right)^{-1}.\end{aligned}$$

⌟

SA-5.5 Control Function: Linear Case (2SLS)

Loosely speaking, control functions are special covariates that can help to eliminate endogeneity issues when added to the estimation problem. The control function relies on a first step estimator and excluded instruments. See Wooldridge (2015) for a recent review.

Due to its popularity in applied work, we will focus on the 2SLS estimator in this subsection, framed as a linear control function approach (the next subsection discusses the non-linear case). We illustrate how overfitting the first step estimate leads to bias in this context. Note that in the "many instruments" literature, it is assumed that $k/n \to C < 1$, and the 2SLS estimator is inconsistent. Here we assume $k = O(\sqrt{n})$, and the 2SLS estimator is consistent, while the distributional approximation is invalid. The result obtained in this section also sheds light on why the JIVE proposed by Imbens et al. (1999) is able to remove the bias, where the special linear structure is key. The next subsection will be devoted to the control function approach in a non-linear setting, where in order to remove the first order bias the jackknife bias correction technique proposed in Section SA-6 is needed, because using JIVE will not suffice.

Consider a simple regression problem with one endogenous regressor X_i and no intercept,

$$Y_i = X_i\theta_0 + u_i,$$

and an auxiliary regression

$$X_i = \mathbf{Z}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i|\mathbf{Z}_i] = 0.$$

We denote $\mu_i = \mathbf{Z}_i^\mathsf{T}\boldsymbol{\beta}$; there is no "misspecification error" for simplicity. To identify the parameter θ_0, we assume $\mathbb{E}[\mu_i^2] \neq 0$ and $\mathbb{E}[u_i|\mathbf{z}_i] = 0$, and regularity conditions such as other moment conditions. The problem can be framed as a control function approach, where the first step residual is plugged-in as an additional regressor in the second step (i.e., a control function approach). Numerically, it is equivalent to the 2SLS approach:

$$\mathbb{E}\left[\mu_i(Y_i - X_i\theta)\right] = 0 \qquad \Leftrightarrow \qquad \theta = \theta_0.$$

Equivalently,

$$\mathbf{w}_i = [Y_i, X_i]^\mathsf{T},\ r_i = X_i,\ \mathbf{z}_i = \mathbf{Z}_i, \qquad \mathbf{m}(\mathbf{w}_i, \mu_i, \theta) = \mu_i(Y_i - X_i\theta).$$

Remark (Using conditional expectation)**.** The above framework encompasses an important case, where some raw instruments $\tilde{\mathbf{Z}}_i \in \mathbb{R}^{d_z}$ are available, which are assumed to be strongly exogenous: $\tilde{\mathbf{Z}}_i \perp\!\!\!\perp u_i$. Of course, it is still possible to project the endogenous regressor X_i onto $\tilde{\mathbf{Z}}_i$ linearly, and the problem will be parametric. On the other hand, it seems natural to exploit the independence to improve efficiency. That is, a function $\mu(\tilde{\mathbf{Z}}_i)$ is found, which explains most of the variation in X_i. It is easy to see that $\mu(\tilde{\mathbf{Z}}_i)$ is the conditional expectation of X_i given $\mathbf{Z}_i$. This is particularly relevant if the endogenous regressor is binary. In this case, $\mathbf{Z}_i$ will be a series expansion of $\tilde{\mathbf{Z}}_i$, and μ_i is essentially a nuisance functional parameter. This is also relevant if $\tilde{\mathbf{Z}}_i$ are categorical. Then the average of X_i in each cell is computed, and depending on the nature of $\tilde{\mathbf{Z}}_i$, the number of cells can be nontrivial compared with the sample size, and the bias could be a serious concern. (Although asymptotically it is a parametric problem since the number of cells is assumed to be fixed.) ⌟

Due to the linear structure, the estimator is

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_{\texttt{2SLS}} - \theta_0\right) &= \left(\frac{1}{n}\sum_i \hat{\mu}_i X_i\right)^{-1} \frac{1}{\sqrt{n}}\sum_i \hat{\mu}_i u_i \\
&= \left(\frac{1}{n}\sum_i \mu_i X_i\right)^{-1} \frac{1}{\sqrt{n}}\sum_i \hat{\mu}_i u_i + o_{\mathbb{P}}(1) \\
&= \left(\frac{1}{n}\sum_i \mu_i X_i\right)^{-1} \frac{1}{\sqrt{n}}\left[\sum_i \mu_i u_i + \sum_i (\hat{\mu}_i - \mu_i) u_i\right] + o_{\mathbb{P}}(1),
\end{aligned}$$

where Assumption A.2(1) is used to justify the second line. An alternative estimator is the JIVE proposed by Imbens et al. (1999), which modifies the first step slightly: instead of using $\hat{\mu}$, the

JIVE uses a leave-one-out version $\hat{\mu}_i^{(i)} = \frac{\hat{\mu}_i}{1-\pi_{ii}} - \frac{\pi_{ii}X_i}{1-\pi_{ii}}$, which gives

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_{\texttt{JIVE}} - \theta_0\right) &= \left(\frac{1}{n}\sum_i \hat{\mu}_i^{(i)} X_i\right)^{-1} \frac{1}{\sqrt{n}}\sum_i \hat{\mu}_i^{(i)} u_i \\ &= \left(\frac{1}{n}\sum_i \mu_i X_i\right)^{-1} \frac{1}{\sqrt{n}}\left[\sum_i \mu_i u_i + \sum_i \left(\frac{\hat{\mu}_i}{1-\pi_{ii}} - \mu_i - \frac{\pi_{ii}X_i}{1-\pi_{ii}}\right) u_i\right] + o_{\mathbb{P}}(1).\end{aligned}$$

Using previous results, it is easy to show that the bias of the 2SLS estimator is

$$\mathcal{B}_{\texttt{2SLS}} = \frac{1}{\sqrt{n}\mathbb{E}[\mu_i^2]}\sum_i \mathbb{E}\left[(\hat{\mu}_i - \mu_i)u_i|\mathbf{Z}_i\right] = \frac{1}{\sqrt{n}\mathbb{E}[\mu_i^2]}\sum_i \mathbb{E}\left[u_i\varepsilon_i|\mathbf{Z}_i\right]\cdot\pi_{ii} = O_{\mathbb{P}}\left(\frac{k}{\sqrt{n}}\right),$$

provided that $\mathbb{E}[\varepsilon_i^4|\mathbf{Z}_i]$ and $\mathbb{E}[u_i^2\varepsilon_i^2|\mathbf{Z}_i]$ are bounded. On the other hand, the JIVE has the bias (provided that $\max_{1\le i\le n}(1-\pi_{ii})^{-1} = O_{\mathbb{P}}(1)$, a necessary condition to "leave-one-out" in the first step):

$$\begin{aligned}\mathcal{B}_{\texttt{JIVE}} &= \frac{1}{\sqrt{n}\mathbb{E}[\mu_i^2]}\sum_i \mathbb{E}\left[\left(\frac{\hat{\mu}_i}{1-\pi_{ii}} - \mu_i - \frac{\pi_{ii}X_i}{1-\pi_{ii}}\right) u_i \middle| \mathbf{Z}_i\right] \\ &= \frac{1}{\sqrt{n}\mathbb{E}[\mu_i^2]}\sum_i \left(\frac{\mathbb{E}[u_i\varepsilon_i|\mathbf{Z}_i]\pi_{ii}}{1-\pi_{ii}} - \frac{\mathbb{E}[u_i\varepsilon_i|\mathbf{Z}_i]\pi_{ii}}{1-\pi_{ii}}\right) = 0,\end{aligned}$$

which shows why the JIVE is able to remove the first order bias.

The above result is not surprising: since the estimating equation is linear in the unobserved quantity μ_i, only the linear bias term (i.e. $b_{1,i}$) is non-zero. The linear bias term is essentially a leave-in bias, hence using a leave-one-out estimator for the first step successfully removes the bias. In the next subsection, we consider the control function approach in a non-linear context. There, the estimating equation will depend on the unobserved quantity μ_i linearly and quadratically, and simply leaving-one-out in the first step (i.e. the JIVE) will not suffice.

SA-5.6 Control Function: Nonlinear Case

To illustrate why the JIVE fails to correct the many instruments bias in a nonlinear setting, we consider the model of Wooldridge (2015):

$$\begin{aligned}Y_i &= \mathbb{1}\left[X_i\cdot\delta_0 \ge u_i\right] \\ X_i &= \mathbf{Z}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i,\end{aligned}$$

where $(u_i, \varepsilon_i) \perp\!\!\!\perp \mathbf{Z}_i$, and has a bivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$. Then, the estimating equation is based on the following conditional expectation:

$$\mathbb{E}[Y_i|X_i, \mathbf{Z}_i] = \mathbb{P}\left[Y_i = 1|X_i, \mathbf{Z}_i\right] = \mathbb{P}\left[u_i \le X_i\delta_0|X_i, \varepsilon_i\right]$$

$$= \Phi\left(X_i\tilde{\delta}_0 - (X_i - \mathbf{Z}_i^\mathsf{T}\boldsymbol{\beta})\gamma_0\right),$$

where Φ is the standard normal c.d.f.,

$$\tilde{\delta}_0 = \delta_0\left(\sigma_{uu} - \frac{\sigma_{u\varepsilon}^2}{\sigma_{\varepsilon\varepsilon}}\right)^{-1/2}, \qquad \gamma_0 = \frac{\sigma_{u\varepsilon}}{\sigma_{\varepsilon\varepsilon}}\left(\sigma_{uu} - \frac{\sigma_{u\varepsilon}^2}{\sigma_{\varepsilon\varepsilon}}\right)^{-1/2},$$

and $\sigma_{u\varepsilon} = \mathbb{E}[u_i\varepsilon]$, $\sigma_{uu} = \mathbb{E}[u_i^2]$ and $\sigma_{\varepsilon\varepsilon} = \mathbb{E}[\varepsilon_i^2]$.

To show the results in a more general context, let $\mu_i = \mathbf{Z}_i^\mathsf{T}\boldsymbol{\beta}$, $\boldsymbol{\theta}_0 = [\tilde{\delta}_0, -\gamma_0]^\mathsf{T}$, and we consider

$$\begin{aligned}
&\mathbf{w}_i = [Y_i, X_i]^\mathsf{T},\ r_i = X_i,\ \mathbf{z}_i = \mathbf{Z}_i \\
&\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = \begin{bmatrix} X_i \\ X_i - \mu_i \end{bmatrix} L'\Big([X_i, X_i - \mu_i]\boldsymbol{\theta}_0\Big)\Big(Y_i - L([X_i, X_i - \mu_i]\boldsymbol{\theta}_0)\Big),
\end{aligned}$$

where L is some prespecified link function. To save notation, let $\mathbf{X}_i = [X_i, X_i - \mu_i]^\mathsf{T}$, then

$$\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)] = \mathbb{E}\left[\mathbf{X}_i L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big)\right] = \mathbf{0}, \tag{E.28}$$

which is essentially the estimating equation for nonlinear least squares. Other exogenous regressors or nonlinear transformations of $X_i - \mu_i$ in $\mathbf{X}_i$ can also be included, which would not change our main conclusion.

Assume $\boldsymbol{\theta}_0$ is identified, which in turn requires that the control function $\varepsilon_i = X_i - \mu_i$ is not degenerate nor perfectly collinear with X_i, and the link function is chosen so that $\mathbb{E}[Y_i|X_i, \mathbf{Z}_i] = L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)$. Then, under standard regularity conditions,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\boldsymbol{\Sigma}_0^{-1}\frac{1}{\sqrt{n}}\sum_i \hat{\mathbf{X}}_i L'(\hat{\mathbf{X}}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\hat{\mathbf{X}}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big) + o_\mathbb{P}(1),$$

with

$$\hat{\mathbf{X}}_i = \begin{bmatrix} X_i \\ X_i - \hat{\mu}_i \end{bmatrix}, \qquad \boldsymbol{\Sigma}_0 = \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\mathsf{T} L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2\right],$$

which is highly nonlinear in the generate regressor $\hat{\mu}_i$.

We summarize the assumptions for this model in the following, in addition to other regularity conditions provided in Section SA-1.

Assumption (Control Function)**.**

A.CF(1) $\boldsymbol{\theta}_0$ is the unique root of the estimating equation (E.28), with some known link function L such that $\mathbb{E}[Y_i|X_i, \mathbf{Z}_i] = L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)$.

A.CF(2) $\varepsilon_i \perp\!\!\!\perp \mathbf{Z}_i$. ⌟

Technically, we do not need to assume L is the conditional expectation of Y_i, neither the independence assumption A.CF(2), as long as (E.28) is taken as the estimating equation and $\boldsymbol{\theta}_0$ is

defined as the parameter of interest. Of course, by dropping those assumptions, $\boldsymbol{\theta}_0$ no longer has a structural interpretation.

Proposition SA.12 (Control Function)**.**
Suppose the assumptions of Theorem SA.6 and CF holds. Then, $\hat{\boldsymbol{\theta}}$ is consistent, and admits the following representation:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{1}{\sqrt{n}}\boldsymbol{\mathcal{B}}\right) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_{\mathbb{P}}(1),$$

where

$$\begin{aligned}
\boldsymbol{\mathcal{B}} &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left[\sum_i \mathbf{b}_{1,i}\cdot\pi_{ii} + \sum_{i,j}\mathbf{b}_{2,ij}\cdot\pi_{ij}^2\right] \\
\bar{\boldsymbol{\Psi}}_1 &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\sum_i \mathbf{X}_i L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big) \\
\bar{\boldsymbol{\Psi}}_2 &= \boldsymbol{\Sigma}_0 \frac{-1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}[\gamma_0 \mathbf{X}_j L'(\mathbf{X}_j^\mathsf{T}\boldsymbol{\theta}_0)^2 | \mathbf{Z}_j]\cdot\pi_{ij}\right)\varepsilon_i,
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{b}_{1,i} &= -\mathbb{E}[\gamma_0 \mathbf{X}_i L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2\varepsilon_i|\mathbf{Z}_i] \\
\mathbf{b}_{2,ij} &= \mathbb{E}\left[\frac{\gamma_0}{2}\mathbf{e}_2 L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2\varepsilon_j^2 - \frac{\gamma_0^2}{2}\mathbf{X}_i L''(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\varepsilon_j^2 - \gamma_0^2\mathbf{X}_i L''(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\varepsilon_j^2 \,\middle|\, \mathbf{Z}_i, \mathbf{Z}_j\right] \\
\boldsymbol{\Sigma}_0 &= \left(\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\mathsf{T} L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2\right]\right)^{-1}.
\end{aligned}$$

If the regressor X_i is in fact exogenous, then $\sigma_{u\varepsilon} = 0$, which means $\gamma_0 = 0$. In this case, the two bias terms will be zero, and the first step has no contribution to the asymptotic variance either. This is not surprising, since then the generated regressor is redundant. In general, however, neither the bias ($\mathbf{b}_{1,i}$ and $\mathbf{b}_{2,ij}$) nor the variance contribution term ($\boldsymbol{\sigma}_2$) will be zero. (Recall that $\mathbf{X}_i$ contains both X_i and ε_i, hence is correlated with ε_i and is not mean zero.)

Due to the presence of the second bias term, $\mathbf{b}_{2,ij}$, the JIVE will not be effective in removing the bias. This is a natural consequence in non-linear models.

SA-5.7 Production Function Estimation

This is the other example included in the main paper. We consider estimation of production functions, following the setup of Olley and Pakes (1996). We first review the setup. Denote by $Y_{i,t}$ the (log) production of firm i at time t, where the production function is taken to be of Cobb-Douglas form with four factors: labor input $L_{i,t}$, capital input $K_{i,t}$, the effect of aging on production

$A_{i,t}$, and a productivity factor $W_{i,t}$. Specifically, we define

$$Y_{i,t} = \beta_L L_{i,t} + \beta_K K_{i,t} + \beta_A A_{i,t} + W_{i,t} + U_{i,t},$$

where the error term $U_{i,t}$ is either measurement error or shock that is unpredictable with time-t information, and has zero conditional mean. Given that the productivity factor is unobserved, the above equation cannot be used directly to estimate the production function.

An investment decision $I_{i,t}$ is based on the productivity factor, hence under some identification assumptions, it is possible to write $W_{i,t} = h_t(I_{i,t}, K_{i,t}, A_{i,t})$, for some unknown and time-dependent function h_t. Therefore, the observed output $Y_{i,t}$ becomes

$$Y_{i,t} = \beta_L L_{i,t} + \phi_t(I_{i,t}, K_{i,t}, A_{i,t}) + U_{i,t}, \qquad \phi_t(I_{i,t}, K_{i,t}, A_{i,t}) = \beta_K K_{i,t} + \beta_A A_{i,t} + h_t(I_{i,t}, K_{i,t}, A_{i,t}).$$

We use $\phi_{i,t} = \phi_t(I_{i,t}, K_{i,t}, A_{i,t})$ whenever possible to save notation. The above display can be used to estimate the labor share β_L, but not β_K or β_A because $h_t(\cdot)$ is unknown.

Taking conditional expectation of $W_{i,t+1}$ on time-t information, and assuming that firm i survives at $t+1$, we have (note the difference in time indexes):

$$\begin{aligned}
Y_{i,t+1} - \beta_L L_{i,t+1} &= \beta_K K_{i,t+1} + \beta_A A_{i,t+1} + g(P_{i,t},\ W_{i,t}) + V_{i,t+1} + U_{i,t+1} \\
&= \beta_K K_{i,t+1} + \beta_A A_{i,t+1} + g(P_{i,t},\ h_t(I_{i,t}, K_{i,t}, A_{i,t})) + V_{i,t+1} + U_{i,t+1} \\
&= \beta_K K_{i,t+1} + \beta_A A_{i,t+1} + g(P_{i,t},\ \phi_{i,t} - \beta_K K_{i,t} - \beta_A A_{i,t}) + V_{i,t+1} + U_{i,t+1},
\end{aligned}$$

where the new error term $V_{i,t+1}$ is the residual from the conditional expectation decomposition of $W_{i,t+1}$, and $P_{i,t}$ is the survival rate,

$$P_{i,t} = \mathbb{P}[\text{firm } i \text{ remains in business at time } t+1 \mid I_{i,t}, A_{i,t}, K_{i,t}] = \mathbb{P}[\chi_{i,t+1} = 1 \mid I_{i,t}, A_{i,t}, K_{i,t}],$$

with $\chi_{i,t}$ the indicator of whether firm i is in business at time t. The residual $U_{i,t+1}$ is orthogonal to time-$t+1$ information, while the residual $V_{i,t+1}$ is obtained from the expectation conditional on time-t information, hence it is *not* orthogonal to information at $t+1$. For this reason, it can be correlated with labor input decision, $L_{i,t+1}$. The term corresponding to labor $\beta_L L_{i,t+1}$ has been moved to LHS for this reason. On the other hand, neither capital nor aging (i.e. $K_{i,t+1}$ and $A_{i,t+1}$) has contemporaneous correlation with the error terms, since they are both "pre-determined".

The three parameters $(\beta_L, \beta_K, \beta_A)'$ are estimated as follows, assuming there are only two time periods $t \in \{t_1, t_2\}$ for simplicity. The labor share β_L is estimated, together with ϕ_{i,t_1}, in a first step with time t_1 data by a partially linear regression: regressing Y_{i,t_1} on L_{i,t_1} and a series expansion of $[I_{i,t_1}, K_{i,t_1}, A_{i,t_1}]^\mathsf{T}$ to obtain $\hat{\beta}_L$ and $\hat{\phi}_{i,t_1}$. Mapping to our generic notation, we define:

$$\begin{aligned}
&r_{1i} = Y_{i,t_1}, \quad z_{11i} = L_{i,t_1}, \quad \mathbf{z}_{12i} = \text{series expansion of } [I_{i,t_1}, K_{i,t_1}, A_{i,t_1}]^\mathsf{T}, \quad \mathbf{z}_{1i} = [z_{11i}, \mathbf{z}_{12i}]^\mathsf{T}, \\
&\nu_{1i} = \beta_L L_{i,t_1} + \mu_{1i} = \beta_L L_{i,t_1} + \phi_t(I_{i,t_1}, K_{i,t_1}, A_{i,t_1}), \quad \varepsilon_{1i} = U_{i,t_1}.
\end{aligned}$$

Another first step is needed to estimate P_{i,t_1}, the survival rate of firm i. We regress/project the indicator of survival χ_{i,t_2} on a series expansion of $[I_{i,t_1}, K_{i,t_1}, A_{i,t_1}]^\mathsf{T}$, and we denote the estimate by $\hat{P}_{i,t_1}$. Mapping to our generic notation, we define:

$$r_{2i} = \chi_{i,t_2}, \quad \mathbf{z}_{2i} = \text{series expansion of } [I_{i,t_1}, K_{i,t_1}, A_{i,t_1}]^\mathsf{T}, \quad \mu_{2i} = P_{i,t_1}, \quad \varepsilon_{2i} = \chi_{i,t_2} - P_{i,t_1}.$$

Finally, assuming that the function $g(\cdot)$ is known up to a finite dimensional parameter $\boldsymbol{\lambda}$, we estimate β_K and β_A in a second step as follows:

$$\underset{\beta_K,\beta_A,\boldsymbol{\lambda}}{\operatorname{argmin}} \frac{1}{n}\sum_i \Big[Y_{i,t_2} - \hat{\beta}_L L_{i,t_2} - \beta_K K_{i,t_2} - \beta_A A_{i,t_2} - g(\hat{P}_{i,t_1}, \hat{\phi}_{i,t_1} - \beta_K K_{i,t_1} - \beta_A A_{i,t_1}, \boldsymbol{\lambda})\Big]^2,$$

which is a standard nonlinear least squares problem. Three quantities are estimated prior to this second step: the labor share β_L and ϕ_{i,t_1} which are jointly estimated in a partially linear first step, and P_{i,t_1} as linear projection in another first step.

It is clear that all our results apply to this example, with the two generalizations proposed in Sections SA-4.1 and SA-4.2. Note that for the two unknowns, ν_{1i} and μ_{2i}, different projection matrices are used. However, we can treat L_{i,t_1} a redundant regressor for estimating P_{i,t_1}. Let $\mathbf{Z}$ be the matrix formed by stacking L_{i,t_1} and series expansion of $[I_{i,t_1}, K_{i,t_1}, A_{i,t_1}]$, and π_{ij} be an element of the projection matrix constructed with $\mathbf{Z}$. Finally, we define the parameter $\boldsymbol{\theta} = [\beta_K, \beta_A, \boldsymbol{\lambda}^\mathsf{T}]^\mathsf{T}$, and the sample moment condition

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\sum_i \mathbf{m}(\mathbf{w}_i,\ \hat{\mu}_{1i},\ \hat{\mu}_{2i},\ \hat{\gamma}, \boldsymbol{\theta}) = \frac{1}{n}\sum_i \mathbf{m}(\mathbf{w}_i,\ \hat{\nu}_{1i} - z_{11i}\hat{\gamma},\ \hat{\mu}_{2i},\ \hat{\gamma}, \boldsymbol{\theta}) \\
&= \frac{1}{n}\sum_i \begin{bmatrix} K_{i,t_1} g_2(\hat{P}_{i,t_1}, \hat{\phi}_{i,t_1} - \beta_K K_{i,t_1} - \beta_A A_{i,t_1}, \boldsymbol{\lambda}) - K_{i,t_2} \\ A_{i,t_2} g_2(\hat{P}_{i,t_1}, \hat{\phi}_{i,t_1} - \beta_K K_{i,t_1} - \beta_A A_{i,t_1}, \boldsymbol{\lambda}) - A_{i,t_2} \\ -\mathbf{g}_3(\hat{P}_{i,t_1}, \hat{\phi}_{i,t_1} - \beta_K K_{i,t_1} - \beta_A A_{i,t_1}, \boldsymbol{\lambda}) \end{bmatrix} \\
&\quad \cdot \Big[Y_{i,t_2} - \hat{\beta}_L L_{i,t_2} - \beta_K K_{i,t_2} - \beta_A A_{i,t_2} - g(\hat{P}_{i,t_1}, \hat{\phi}_{i,t_1} - \beta_K K_{i,t_1} - \beta_A A_{i,t_1}, \boldsymbol{\lambda})\Big]
\end{aligned}$$

with $\mathbf{w}_i = [Y_{i,t_2}, L_{i,t_2}, K_{i,t_2}, A_{i,t_2}, L_{i,t_1}, K_{i,t_1}, A_{i,t_1}]'$, $\hat{\gamma} = \hat{\beta}_L$, g_ℓ denoting the derivative of g with respect to its ℓ-th argument, and analogously for higher order derivatives.

Proposition SA.13 (Production Function)**.**
Suppose the assumptions of Theorem SA.5, the additional regularity conditions discussed in Sections SA-4.1 and SA-4.2, and $\sum_i \pi_{ii}^2 = o_\mathbb{P}(k)$ hold. Then, $\hat{\boldsymbol{\theta}}$ is consistent, and admits the following representation:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{1}{\sqrt{n}}\boldsymbol{\mathcal{B}}\right) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_\mathbb{P}(1),$$

where

$$\boldsymbol{\mathcal{B}} = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \left[\sum_i \Big(\mathbf{b}_{1,1,i} + \mathbf{b}_{1,2,i} \Big) \pi_{ii} + \sum_{i,j} \Big(\mathbf{b}_{2,11,ij} + \mathbf{b}_{2,22,ij} + \mathbf{b}_{2,12,ij} \Big) \pi_{ij}^2 \right],$$

$$\bar{\boldsymbol{\Psi}}_1 = \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_i \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} \Big(V_{i,t_2} + U_{i,t_2} \Big),$$

$$\bar{\boldsymbol{\Psi}}_2 = -\frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_i \left\{ \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{2,i,t_1} U_{i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{1,i,t_1} \Big(\chi_{i,t_2} - P_{i,t_1} \Big) \right\}$$
$$+ \frac{1}{\sqrt{n}} \frac{1}{\mathbb{E}\mathbb{V}[L_{i,t_1} | (I,K,A)_{i,t_1}]} \boldsymbol{\Sigma}_0 \boldsymbol{\Xi}_0 \sum_i \Big(L_{i,t_1} - \mathbb{E}[L_{i,t_1} | (I,K,A)_{i,t_1}] \Big) U_{i,t_1},$$

and

$$\mathbf{b}_{1,1,i} = \begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1} \end{bmatrix} \mathbb{C}\mathrm{ov}\Big[V_{i,t_2}, U_{i,t_1} \Big| (L,I,K,A)_{i,t_1} \Big]$$

$$\mathbf{b}_{1,2,i} = \begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1} \end{bmatrix} \mathbb{C}\mathrm{ov}\Big[V_{i,t_2}, \chi_{i,t_2} \Big| (L,I,K,A)_{i,t_1} \Big]$$

$$\mathbf{b}_{2,11,ij} = -\frac{1}{2} \left\{ 2 \begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1} \end{bmatrix} g_{2,i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{22,i,t_1} \right\} \mathbb{V}\Big[U_{j,t_1} \Big| (L,I,K,A)_{j,t_1} \Big],$$

$$\mathbf{b}_{2,22,ij} = -\frac{1}{2} \left\{ 2 \begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1} \end{bmatrix} g_{1,i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{11,i,t_1} \right\} \mathbb{V}\Big[\chi_{j,t_2} \Big| (L,I,K,A)_{j,t_1} \Big],$$

$$\mathbf{b}_{2,12,ij} = -\left\{ -\begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1} \end{bmatrix} g_{1,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1} \end{bmatrix} g_{2,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{12,i,t_1} \right\}$$
$$\mathbb{C}\mathrm{ov}\Big[U_{j,t_1}, \chi_{j,t_2} \Big| (L,I,K,A)_{j,t_1} \Big].$$

⌟

We do not provide formulas for $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Xi}_0$ to save space, but they follow formally from our general results discussed in Sections SA-3 and SA-4. Furthermore, we made the additional assumption that $\sum_i \pi_{ii}^2 = o_{\mathbb{P}}(k)$ to simplify the bias formula only; the result remains true without this assumption, albeit the biases become even more cumbersome.

SA-5.8 Conditional Moment Restrictions

The 2SLS estimator is closely related to another class of problems defined by conditional moment restrictions. Let $\mathbb{E}[e(Y_i, X_i, \theta_0)|\mathbf{Z}_i] = 0$, where Y_i is the outcome variable, X_i is the endogenous regressor, and $\mathbf{Z}_i$ are excluded instruments (or transformations thereof), and $e(\cdot)$ is known up to the finite dimensional parameter θ_0. An unconditional moment restriction is $\mathbb{E}[g(\mathbf{Z}_i)e(Y_i, X_i, \theta_0)] = 0$, for some function $g(\mathbf{Z}_i)$, provided the parameter of interest θ_0 remains identifiable.

One particular choice is the following:

$$\mathbb{E}[\mu_i e(Y_i, X_i, \theta_0)] = 0, \qquad \mu_i = \mathbf{Z}_i^{\mathsf{T}}\boldsymbol{\beta}, \tag{E.29}$$

and $\boldsymbol{\beta}$ is the (population) regression coefficient of X_i on $\mathbf{Z}_i$. This reduces to the 2SLS estimator already discussed above when $e(Y_i, X_i, \theta_0) = Y_i - X_i\theta_0$. And, in fact, this choice will be optimal in the sense of Wooldridge (2010, Section 14.4.3) under (conditional) homoskedasticity. Nevertheless, we take the estimating equation (E.29) as given, and investigate how the first step estimate affects the asymptotic distribution of $\hat{\theta}$. Since the estimator (or estimating equation) is linear in the first step estimator $\hat{\mu}_i$, it is easy to show that $\sqrt{n}(\hat{\theta} - \theta_0)$ has the following first order bias:

$$\mathcal{B} = -\left(\mathbb{E}\left[\mu_i \frac{\partial}{\partial\theta} e(Y_i, X_i, \theta_0)\right]\right)^{-1} \frac{1}{\sqrt{n}} \sum_i \mathbb{E}\left[e(Y_i, X_i, \theta_0)\varepsilon_i|\mathbf{Z}_i\right] \pi_{ii} = O_{\mathbb{P}}(k/\sqrt{n}),$$

under regularity conditions. The same arguments made previously for the 2SLS estimator apply here: the bias is essentially a leave-in bias, hence a simple JIVE is effective for bias correction.

The choice of instrument in (E.29) is arbitrary and, in general, not optimal. A more interesting behavior arises when the optimal instrument, under possibly conditional heteroskedasticity, is used. This optimal instrument is

$$\frac{\mu_{1i}}{\mu_{2i}} = \frac{\mathbb{E}[\partial e(Y_i, X_i, \theta_0)/\partial\theta|\mathbf{Z}_i]}{\mathbb{V}[e(Y_i, X_i, \theta_0)|\mathbf{Z}_i]},$$

which requires estimating two unknown quantities: μ_{1i} and μ_{2i}. See Section SA-4.1 for this generalization. Our results also apply to this case, though characterizing the leading, many covariates bias is very cumbersome.

Depending on the specific context, the JIVE may or may not be effective for bias correction. To be more concrete, consider first the homoskedastic case, where the optimal instrument reduces to $\mu_{1i} = \mathbb{E}[\partial e(Y_i, X_i, \theta_0)/\partial\theta|\mathbf{Z}_i]$, which can still be estimated by a linear projection. In this case, the JIVE is effective since the estimating equation is linear in the (unknown) instrument. However, consider now the general (conditional) heteroskedastic case, where the instrument is the ratio of two unknown functions, and the denominator is obtained by regressing $e(Y_i, X_i, \theta_0)^2$ on $\mathbf{Z}_i$. The estimating equation is now nonlinear in μ_{2i}, and therefore the leading many covariates bias is no longer a leave-in bias only, which implies that the JIVE is no longer effective in removing this bias. Our generic fully data-driven results do apply, and our proposed jackknife bias-correction and bootstrap-based inference can be used directly in this case.

SA-6 The Jackknife

We show that the jackknife is able to estimate consistently the many instrument bias and the asymptotic variance, even when many instruments are used (i.e., $k = O(\sqrt{n})$). We first describe the data-driven, fully automatic algorithm.

Algorithm SA.1 (Jackknife)**.**

Step 1. For each observation $j = 1, 2, \ldots, n$ estimate μ_i without using the j-th observation, which we denote by $\hat{\mu}_i^{(j)}$, and compute the leave-j-out estimator by solving:

$$\hat{\boldsymbol{\theta}}^{(j)} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left| \boldsymbol{\Omega}_n^{1/2} \sum_{i, i \neq j} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{(j)}, \boldsymbol{\theta}) \right|,$$

where, taking the estimator as a black box, this step simply requires to delete the j-th row from the data matrix because

$$\hat{\mu}_i^{(j)} = \hat{\mu}_i + \frac{\pi_{ij}}{1 - \pi_{jj}} \Big(\hat{\mu}_j - r_j \Big), \qquad 1 \leq i \leq n.$$

Define $\hat{\boldsymbol{\theta}}^{(\cdot)} = \frac{1}{n} \sum_j \hat{\boldsymbol{\theta}}^{(j)}$.

Step 2. The jackknife bias estimator is defined as

$$\hat{\boldsymbol{\mathcal{B}}} = (n-1) \cdot \sqrt{n} \left(\hat{\boldsymbol{\theta}}^{(\cdot)} - \hat{\boldsymbol{\theta}} \right) = \frac{n-1}{n} \sum_j \sqrt{n} \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}} \right), \tag{E.30}$$

and the bias corrected estimator is $\hat{\boldsymbol{\theta}}_{\mathsf{bc}} = \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mathcal{B}}}/\sqrt{n}$.

Step 3. The jackknife variance estimator is

$$\hat{\boldsymbol{\mathcal{V}}} = (n-1) \sum_j \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)} \right) \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)} \right)^{\mathsf{T}}. \tag{E.31}$$

⌟

To match the notation used in the main paper, note that $\hat{\mathscr{B}} = \hat{\boldsymbol{\mathcal{B}}}/\sqrt{n}$, and therefore $\hat{\boldsymbol{\theta}}_{\mathsf{bc}} = \hat{\boldsymbol{\theta}} - \hat{\mathscr{B}}$. Similarly, $\hat{\mathscr{V}} = \hat{\boldsymbol{\mathcal{V}}}/n$. The reason we introduce $\hat{\boldsymbol{\mathcal{B}}}$ and $\hat{\boldsymbol{\mathcal{V}}}$ is that they are asymptotically non-vanishing, under the assumption $k \propto \sqrt{n}$.

To show the validity of the jackknife, we impose the following additional assumption.

Assumption A.3 (Design Balance)**.**

A.3(1) $\sum_i \pi_{ii}^2 = o_{\mathbb{P}}(k)$;

A.3(2) $\max_{1 \leq i \leq n} 1/(1 - \pi_{ii}) = O_{\mathbb{P}}(1)$. ⌟

Proposition SA.14 (Jackknife Validity).
Assume A.1, A.2 and A.3 hold, and $k = O(\sqrt{n})$. Then the jackknife bias correction estimate (E.30) and variance estimate (E.31) are consistent:

$$\boldsymbol{\mathcal{B}} - \hat{\boldsymbol{\mathcal{B}}} = o_{\mathbb{P}}(1), \qquad \mathbb{V}[\mathbb{E}[\bar{\boldsymbol{\Psi}}_1|\mathbf{Z}]] + \mathbb{V}\left[\bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 \middle| \mathbf{Z}\right] - \hat{\boldsymbol{\mathcal{V}}} = o_{\mathbb{P}}(1).$$

⌟

SA-7 The Bootstrap

Although bias correction will not affect the variability of the estimator asymptotically, it is likely to have impact in finite samples. One remedy is to embed the jackknife bias correction into nonparametric bootstraps. To be more specific, one first samples with replacement, and then obtains bias corrected estimator from the bootstrap sample. For nonlinear estimation problems, however, the nonparametric bootstrap may not be appealing, since numerical procedures can fail to converge for the bootstrap data.

In this section we propose a new bootstrap procedure, which combines the wild bootstrap and the multiplier bootstrap. Two separate aspects of the bootstrap will be discussed. First we show that the bootstrap can be used to estimate the bias, and provides valid distributional approximation. Second, the jackknife can be embedded into the bootstrap, which allows one to bootstrap the studentised and bias-corrected statistic, and yields better distributional approximation after bias correction.

SA-7.1 Large Sample Properties

First we describe the bootstrap procedure without embedding the jackknife. Let $\{e_i^\star\}_{1\le i\le n}$ be i.i.d. bootstrap weights orthogonal to the original data, and have zero mean and unit variance (also finite fourth moment). Then we use the wild bootstrap for the first step. More explicitly,

$$\begin{aligned}\hat{\mu}_i^\star &= \mathbf{z}_i^\mathsf{T}\left(\sum_j \mathbf{z}_j\mathbf{z}_j^\mathsf{T}\right)^- \left(\sum_j \mathbf{z}_j(\hat{\mu}_j + \hat{\varepsilon}_j \cdot e_j^\star)\right) \\ &= \hat{\mu}_i + \mathbf{z}_i^\mathsf{T}\left(\sum_j \mathbf{z}_j\mathbf{z}_j^\mathsf{T}\right)^- \left(\sum_j \mathbf{z}_j\hat{\varepsilon}_j \cdot e_j^\star\right), \qquad \hat{\varepsilon}_j = r_j - \hat{\mu}_j. \end{aligned} \tag{E.32}$$

For the second step, $\hat{\boldsymbol{\theta}}^\star$ solves the following moment condition (called the multiplier bootstrap):

$$\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\right]^\mathsf{T} \boldsymbol{\Omega}_n \left[\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^\star, \hat{\boldsymbol{\theta}}^\star)\cdot(1+e_i^\star)\right] = o_{\mathbb{P}}(1), \tag{E.33}$$

which is the bootstrap analogue to (E.5).

Remark (Vector-Valued $\boldsymbol{\mu}_i$)**.** Nothing changes in the bootstrap procedure when there are multiple unknowns to be estimated in the first step (see Section SA-4.1). To implement the bootstrap, we would like to mention that the same bootstrap weight $e_i^\star$ has to be used for generating $\hat{\mu}_{\ell i}^\star$:

$$\hat{\mu}_{\ell i}^\star = \hat{\mu}_{\ell i} + \mathbf{z}_i^\mathsf{T} \left(\sum_j \mathbf{z}_j \mathbf{z}_j^\mathsf{T} \right)^- \left(\sum_j \mathbf{z}_j \hat{\varepsilon}_{\ell j} \cdot e_j^\star \right), \qquad \hat{\varepsilon}_{\ell j} = r_{\ell j} - \hat{\mu}_{\ell j},$$

for $1 \leq \ell \leq d_\mu$. The second step remains the same. ⌟

The following conditions are useful to establish results using the bootstrap.

Assumption A.4 (Bootstrap)**.**

A.4(1) $\hat{\boldsymbol{\theta}}^\star$ is given by (E.32) and (E.33), and is tight.

A.4(2) $\hat{\mu}_i^\star$ is uniformly consistent: $\max_{1 \leq i \leq n} |\hat{\mu}_i^\star - \hat{\mu}_i| = o_\mathbb{P}(1)$. ⌟

Assumption A.4(2) is the bootstrap analogue of Assumption A.2(1), and can be established with suitable primitive conditions along the lines already discussed in Section SA-2. As an example, we provide the following lemma.

Lemma SA.15.
Suppose Assumption A.2(1) holds and, in addition, (i) $\mathbb{E}[\exp(e_i^{\star 2}/M^2)] < \infty$ for some $M < \infty$ and (ii) $(\max_{1 \leq i \leq n} \varepsilon_i^2)(\max_{1 \leq i \leq n} \pi_{ii}) = o_\mathbb{P}(1/\log(n))$. Then, A.4(2) holds. ⌟

The primitive conditions required in this lemma have an intuitive interpretation. First, (i) concerns minimal requirements on the bootstrap weights, allowing for all type of weights with compact support or with sub-Gaussian tails. Second, (ii) imposes conditions on the tail of the distribution of the residual $\varepsilon_i = r_i - \mu_i$ together with restrictions on the first-step covariates via the statistic $\max_{1 \leq i \leq n} \pi_{ii}$. Since ε_i is not degenerate at 0, (ii) implies $\pi_{ii} = o_\mathbb{P}(1/\log(n))$ but usually a little more will be required because $\max_{1 \leq i \leq n} \varepsilon_i^2 \to_\mathbb{P} \infty$ unless ε_i has bounded support. For example, if the residuals are Gaussian, then $\max_{1 \leq i \leq n} \varepsilon_i^2 \asymp_\mathbb{P} \log n$. The discussion in Section SA-2 can be used to give primitive conditions on $\mathbf{z}_i$ ensuring that (ii) holds.

Now we state a result that is similar to Proposition SA.5

Proposition SA.16 (Asymptotic Representation: Bootstrap)**.**
Assume A.1, A.2 and A.4 hold, and $k = O(\sqrt{n})$. Then

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \frac{\boldsymbol{\mathcal{B}} + \boldsymbol{\mathcal{B}}'}{\sqrt{n}} \right) = \bar{\boldsymbol{\Psi}}_1^\star + \bar{\boldsymbol{\Psi}}_2^\star + o_\mathbb{P}(1),$$

where $\boldsymbol{\mathcal{B}}$ is given in Proposition SA.5, and

$$\boldsymbol{\mathcal{B}}' = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \left[\sum_i \mathbf{b}_{2,ii} \cdot \pi_{ii}^2 \cdot \mathbb{E}[e_i^{\star 3}] \right]$$

$$\bar{\boldsymbol{\Psi}}_1^\star = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \left[\sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \cdot e_i^\star \right] \qquad \bar{\boldsymbol{\Psi}}_2^\star = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \left[\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0) \ |\mathbf{z}_j] \pi_{ij} \right) \varepsilon_i \cdot e_i^\star \right].$$

⌟

Finally we note that without bias, bootstrap consistency can be established easily by appealing to Lindeberg-type CLT arguments, by conditioning on the original data. On the other hand, the bootstrap is able to replicate the many covariates/instruments bias only under the assumption that $\mathcal{B}' = o_{\mathbb{P}}(1)$, which can be achieved by using bootstrap weights $e_i^\star$ with zero third moment.

SA-7.2 Bootstrapping Bias-Corrected Estimators

Section SA-6 proposes the jackknife as a method for bias correction and variance estimation. In particular, it is showed that $\hat{\boldsymbol{\mathcal{B}}}$ is first order equivalent to the $\boldsymbol{\mathcal{B}}$, hence is asymptotically degenerate (i.e. does not contribute to variance). On the other hand, it should be expected that in finite samples, bias correction injects noise, which will affect the performance of distributional approximations.

In this subsection, we combine the bootstrap and the jackknife. More specifically, the jackknife bias correction and variance estimation are embedded into the bootstrap, which makes it possible to bootstrap the bias-corrected and studentised statistic (that is, bootstrap the bias-corrected t-statistic).

Algorithm SA.2 (Bootstrapping Bias-Corrected and Studentised Statistics)**.**

Step 1. Apply Algorithm SA.1 and construct the bias-corrected t-statistic $\mathscr{T} = (\hat{\boldsymbol{\mathcal{V}}}/n)^{-1/2}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \hat{\boldsymbol{\mathcal{B}}}/\sqrt{n}\Big)$.

Step 2. Compute $\hat{\boldsymbol{\theta}}^{\star,(j)}$ as

$$\hat{\boldsymbol{\theta}}^{\star,(j)} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left| \boldsymbol{\Omega}_n^{1/2} \sum_i \Big(e_i^\star + \mathbb{1}[i \neq j] \Big) \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{\star,(j)}, \boldsymbol{\theta}) \right|$$

$$\hat{\boldsymbol{\theta}}^{\star,(\cdot)} = \sum_j (1 + e_j^\star) \hat{\boldsymbol{\theta}}^{\star,(j)} \Big/ \sum_j (1 + e_j^\star),$$

where $\hat{\mu}_i^{\star,(j)}$ is obtained by regressing $r_i^\star$ on $\mathbf{z}_i$, without using the j-th observation. Then

$$\hat{\boldsymbol{\mathcal{B}}}^\star = (n-1)\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}^\star\right), \qquad \hat{\boldsymbol{\mathcal{V}}}^\star = (n-1)\sum_j (1 + e_j^\star)\left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)}\right)\left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)}\right)^{\mathsf{T}}.$$

Then construct $\mathscr{T}^\star = (\hat{\boldsymbol{\mathcal{V}}}^\star/n)^{-1/2}\Big(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n}\Big)$.

Step 3. Repeat the previous step, and use the empirical distribution of $\mathscr{T}^\star$ to approximate that of $\mathscr{T}$. ⌟

In the main paper, we use different scaling:

$$\hat{\mathscr{B}}^\star = (n-1)\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}^\star\right), \qquad \hat{\mathscr{V}}^\star = \frac{n-1}{n}\sum_j (1+e_j^\star)\left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)}\right)\left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)}\right)^{\mathsf{T}},$$

and therefore $\mathscr{T}^\star = (\hat{\mathscr{V}}^\star)^{-1/2}\left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \hat{\mathscr{B}}^\star\right)$.

Remark (Centering the bootstrap distribution)**.** Asymptotically the distribution of $\mathscr{T}^\star$ is centered at the origin, since the bias correction term $\hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n}$ is consistent. In finite samples, this may not be true, and can be problematic. A practical solution is to use $\mathscr{T}^\star = (\hat{\boldsymbol{\mathcal{V}}}^\star/n)^{-1/2}\Big(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n} - \mathbb{E}^\star[\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n}]\Big)$. ⌟

Remark (Failure of naïve jackknife)**.** Employing the jackknife on top of the multiplier bootstrap requires reweighting the bias and variance estimators. This is a generic issue for any bootstrap employed in multiplier form, including the standard nonparametric bootstrap. The "naïve" way of implementing the jackknife under the bootstrap would delete one observation each time in the second step, that is, $\hat{\boldsymbol{\theta}}^{\star,(\ell)} = \operatorname{argmin}_{\boldsymbol{\theta}} \left|\boldsymbol{\Omega}_n^{1/2}\sum_{i=1,i\neq\ell}^n \omega_i^\star \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{\star,(\ell)}, \boldsymbol{\theta})\right|$. This approach does not work in general because the resulting variance estimator is inconsistent. To see this, observe that this naïve jackknife approach (under the multiplier bootstrap distribution) ignores the bootstrap weighting scheme and by deleting observations together with the associated weights, it effectively deletes "blocks of observations", thereby introducing extra variability, which makes the variance estimator inconsistent. ⌟

For the remaining of this section, we consider the properties of the jackknife bias and variance estimator applied to the bootstrapped sample. The techniques we use will be similar to those of Proposition SA.14 and SA.16.

Proposition SA.17 (Jackknife Validity with Bootstrapped Sample)**.**
Assume A.1, A.2, A.3 and A.4 hold, and $k = O(\sqrt{n})$. In addition, assume the bootstrap weights $e_i^\star$ have zero third moment. Then

$$\boldsymbol{\mathcal{B}} + \boldsymbol{\mathcal{B}}' - \hat{\boldsymbol{\mathcal{B}}}^\star = o_{\mathbb{P}}(1), \qquad \mathbb{V}^\star[\bar{\boldsymbol{\Psi}}_1^\star + \bar{\boldsymbol{\Psi}}_2^\star] - \hat{\boldsymbol{\mathcal{V}}}^\star = o_{\mathbb{P}}(1).$$

⌟

SA-8 Numerical Evidence

In this section we provide numerical evidence of the many-covariates bias we found in Section SA-3, and demonstrate the jackknife bias correction technique proposed in Section SA-6. For better

inference, we bootstrap the bias-corrected test statistic (c.f. Section SA-7). We illustrate with both simulation studies and an empirical exercise, in the context of marginal treatment effects (c.f. Section SA-5.4).

SA-8.1 Monte Carlo Experiments

In this section, we consider three sets of simulations for the marginal treatment effects. **DGP 1**: the propensity score is low dimensional and correctly specified, while we add redundant covariates to the first step and see the consequence. **DGP 2**: the propensity score is nonlinear in the covariates and has moderate dimension, and we consider the pseudo true value of the marginal treatment effect corresponding to a linear approximation to the propensity score. Again we add redundant covariates to the first step to increase the dimension. **DGP 3**: the propensity score is nonlinear with low dimension. We consider using a series approximation, hence in the limit, the propensity score will be correctly specified. Therefore we are able to illustrate two sources of biases: bias due to misspecified propensity score (when k is small), and bias due to many covariates (when k is large).

For each data generating process, we use three methods to conduct inference. The first method relies on the bootstrap only, as we showed that the bootstrap is able to approximate the distribution (including the bias due to many covariates). The second method replies on the jackknife only. While the last method utilizes both the jackknife and the bootstrap. In particular, we bootstrap the jackknife bias-corrected t-statistic.

DGP 1. (Table 1–3) Let the potential outcomes be $Y_i(0) = U_{0i}$ and $Y_i(1) = 0.5+U_{1i}$. We assume there are many potential covariates $\mathbf{Z}_i = [1, \{Z_{\ell,i}\}_{1\le\ell\le 199}]$, with $Z_{\ell,i} \sim \text{Uniform}[0, 0.2]$ independent across ℓ and i. To illustrate the bias and size distortion due to many covariates, without being contaminated by misspecified propensity score, the selection equation is assumed to take a very parsimonious form: $T_i = \mathbb{1}\left[0.1 + \sum_{\ell=1}^{4} Z_{\ell,i} \ge V_i\right]$. Finally the error terms are distributed as $V_i|\mathbf{Z}_i \sim \text{Uniform}[0, 1]$, $U_{0i}|\mathbf{Z}_i, V_i \sim \text{Uniform}[-1, 1]$ and $U_{1i}|\mathbf{Z}_i, V_i \sim \text{Uniform}[-0.5, 1.5 - 2V_i]$. Note that we do not have any covariates $\mathbf{X}_i$ here, and the treatment effect heterogeneity and self-selection are captured by the correlation between U_{1i} and V_i. Then $\mathbb{E}[Y_i|P_i = a] = a - \frac{a^2}{2}$ and the MTE is $\tau_{\texttt{MTE}}(a) = 1 - a$. To estimate MTE, set $\mathbf{X}_i = 1$ and $\boldsymbol{\phi}(p) = p^2$, and the second step regression becomes $\hat{\mathbb{E}}[Y_i|P_i] = \hat{\theta}_1 + \hat{\theta}_2 \cdot \hat{P}_i + \hat{\theta}_3 \cdot \hat{P}_i^2$. The estimated MTE is $\hat{\tau}_{\texttt{MTE}}(a) = \hat{\theta}_2 + 2a \cdot \hat{\theta}_3$. In simulation, we consider the normalized quantity $\sqrt{n}\left(\hat{\tau}_{\texttt{MTE}}(a) - \tau_{\texttt{MTE}}(a)\right)$ at $a = 0.5$, with and without bias correction. The sample sizes are $n \in \{1000, 2000\}$, and we use 2000 Monte Carlo repetitions. To estimate the propensity score, we regress T_i on a constant term and $\{Z_{\ell,i}\}$ for $1 \le \ell \le k - 1$, where the number of covariates k ranges from 5 to 200. Note that $k = 5$ corresponds to the most parsimonious model which is correctly specified.

In the tables we illustrate the empirical bias (column "bias"), standard deviation (column "sd"), empirical size of a level-0.1 test (columns "size†" and "size‡"), and length of confidence interval (columns "ci†" and "ci‡"). For the empirical size and CI length, we use two approaches to illustrate the effect of bias correction. The first approach ignores the problem of variance estimation. That

is, instead of using standard errors, the test statistics are constructed by using the oracle standard error (that is, the standard deviation of the estimator obtained from simulation). Results from this approach correspond to columns "size†" and "ci†".

The second approach we take concerns the performance of bias correction in a feasible setting. With the bootstrap, we simply use the empirical distribution to conduct hypothesis testing. And if only the jackknife is used, we rely on the feasible jackknife variance estimator to construct the t-statistic, and the inference is based on normal approximation. Results from the second approach correspond to columns "size‡" and "ci‡".

Table 1 collects the simulation results when only the bootstrap is used. First it is obvious that without bias correction, the asymptotic bias shows up quickly as k increases, which leads to severe size distortion. Interestingly, the finite sample variance shrinks at the same time. Therefore for this particular DGP, incorporating many not only leads to biased estimates, but also gives the illusion that the MTE is estimated precisely. Recall that the $k = 5$ model is correctly specified, therefore the variance there reflects the true variability of the estimator. The bootstrap can partially remove the bias and restore the empirical size closer to its nominal level, as the bootstrap distribution captures the many-covariate bias. In Table 2, we only use the jackknife for bias correction and variance estimation. Compared with the bootstrap, the jackknife performs much better in terms of correcting bias, although the bias correction introduces additional noise in finite samples. Finally in Table 3, we combine the jackknife and the bootstrap, since the jackknife delivers excellent bias correction and the bootstrap is able to take into account the additional variation. One can see that the empirical coverage rate remains well-controlled even with 100 covariates used in the first step.

Although the focus here is inference and the size distortion issue, it is also important to know how bias correction will affect the mean squared error (MSE), a criterion commonly used to evaluate estimators. Recall that the model is correctly specified with five covariates (i.e. $k = 5$), hence it should not be surprising that incorporating bias correction there increases the variability of the estimator and the MSE – although the impact is very small. As more covariates are included, however, the MSE increases rapidly without bias correction, while the MSE of the bias corrected estimator remains relatively stable. Therefore the bias-corrected estimator is not only appealing for inference – it also performs better in terms of MSE when the number of covariates is moderate or large.

DGP 2. (Table 4–6) To illustrate the implications of using many covariates in a more realistic setting, we make some modifications of the previous data generating process. The selection equation now depends on many more covariates, $T_i = \mathbb{1}\left[\Phi\left(0.5\sum_{\ell=1}^{49} Z_{\ell,i} - 12.25\right) \geq V_i\right]$, where Φ is the standard normal c.d.f., and the covariates are i.i.d. uniformly distributed on $[0, 1]$. Since we do not change the joint distribution of the error terms, the marginal treatment effect remains to be $\tau_{\texttt{MTE}}(a) = 1 - a$.

For estimation, we still fit a linear model for the propensity score. By doing so, the propensity score will be misspecified regardless of the number of covariates used, and the true MTE cannot be recovered. On the other hand, this can be understood as estimating a pseudo-true value, which is

defined as the "MTE identified with a linear approximation to the propensity score". In simulations, we center the test statistic at the pseudo-true MTE, rather than the population MTE, which is obtained from a simulation with 50 covariates and very large sample size (the centering is 0.545 when $a = 0.5$).

Since the pseudo-true MTE is obtained by using 50 covariates, there will be misspecification bias when $k < 50$. This is indeed confirmed by the simulation. When the number of included covariates is beyond 50, the models can be regarded as correctly specified for the pseudo-true MTE. With large k, however, the many covariates bias will dominate and lead to severe size distortion without bias correction. The bias-corrected estimator, on the other hand, removes most of the bias and the empirical coverage is very close to the nominal level.

DGP 3. (Table 7–9) In the final set of simulations, series estimation (see the following table for details) is used to estimate a nonlinear propensity score. We center the test statistic by the true MTE, and the misspecification error will decrease (although never disappear) with more covariates used. To be more precise, the selection equation is $T_i = \mathbb{1}\left[\Phi\left(\sum_{\ell=1}^{5} Z_{\ell,i} - 3.5\right) \geq V_i\right]$, which depends nonlinearly on five "raw covariates", uniformly distributed on $[0, 1]$. To fit the model flexibly, we gradually include more interactions and higher-order terms of the raw covariates. Note that when k is small, the bias mainly comes from misspecifying the propensity score, while for large k, the many covariates bias will dominate. This is indeed confirmed by the simulation results (the empirical coverage exhibits inverted-V shape without bias correction). With bias correction, the many covariates bias is much better controlled. Moreover, the two estimators exhibit similar MSEs when k is small, while the bias-corrected estimator has much smaller MSE when k is moderate or large.

Polynomial Basis Expansion.

k	$\mathbf{s}^k(\mathbf{Z}_i)$	k	$\mathbf{s}^k(\mathbf{Z}_i)$
6	1 and $\mathbf{Z}_i$	61	$\mathbf{s}^{56}(\mathbf{Z}_i)$ and $[Z_{1i}^4, Z_{2i}^4, \cdots, Z_{5i}^4]$
11	1, $\mathbf{Z}_i$ and $[Z_{1i}^2, Z_{2i}^2, \cdots, Z_{5i}^2]$	126	$\mathbf{s}^{61}(\mathbf{Z}_i)$ and 4^{th}-order interactions
21	$\mathbf{s}^{11}(\mathbf{Z}_i)$ and 2^{nd}-order interactions	131	$\mathbf{s}^{126}(\mathbf{Z}_i)$ and $[Z_{1i}^5, Z_{2i}^5, \cdots, Z_{5i}^5]$
26	$\mathbf{s}^{21}(\mathbf{Z}_i)$ and $[Z_{1i}^3, Z_{2i}^3, \cdots, Z_{5i}^3]$	252	$\mathbf{s}^{131}(\mathbf{Z}_i)$ and 5^{th}-order interactions
56	$\mathbf{s}^{26}(\mathbf{Z}_i)$ and 3^{rd}-order interactions	257	$\mathbf{s}^{252}(\mathbf{Z}_i)$ and $[Z_{1i}^6, Z_{2i}^6, \cdots, Z_{5i}^6]$

SA-8.2 Empirical Illustration

In this section we report the marginal returns to college education with the data used in Carneiro, Heckman and Vytlacil (2011), estimated by the local instrumental variable approach. Moreover, we illustrate the importance of employing bias correction, and how it affects the estimated treatment effect heterogeneity.

The data consists of a subsample of white males from the 1979 National Longitudinal Survey of Youth (NLSY79), and the sample size is $n = 1,747$. The outcome variable, Y_i, is the log wage in 1991, and the sample is split according to the treatment variable $T_i = 0$ (high school dropouts

and high school graduates), and $T_i = 1$ (with some college education or college graduates). Hence the parameter of interest is the return to college education. The dataset includes covariates on individual and family background information, and four "raw" instrumental variables: presence of four-year college, average tuition, local unemployment and wage rate, measured at age 17 of the survey participants. We follow Carneiro et al. (2011) and normalize the estimates by the difference of average education level between the two groups, so that the estimates are interpreted as return to per year college education. The summary statistics are given in Table 10.

Standard linear regression (OLS) yields point estimate 0.072 (standard error 0.007), and two-stage least squares (2SLS) using the aforementioned instruments yields 0.155 (standard error 0.048). Argued in Heckman and Vytlacil (2005), the 2SLS estimate is hard to interpret in practice (unless the instrument is binary) for two reasons. First it does not provide information on treatment effect heterogeneity, which is crucial for many economic/policy questions. Second, the 2SLS is a complicated weighted average of the marginal treatment effect, which many not reflect the effect of any policy experiment. We employ the local instrumental variable approach to estimate the marginal treatment effect, as well as the bias correction technique we proposed in this paper.

Different sets of covariates are defined in the following, for future reference.

(i) linear and square terms of corrected AFQT score, education of mom, number of siblings, permanent average local unemployment rate and wage rate at age 17;

(ii) indicator of urban residency at age 14;

(iii) cohort dummy variables;

(iv) average local unemployment rate and wage rate in 1991, and linear and square terms of work experience in 1991.

(v) the four raw instruments: presence of four-year college, average local college tuition at age 17, average local unemployment and wage rate at age 17, as well as their interactions with corrected AFQT score, education of mom and number of siblings.

(vi) interactions among corrected AFQT score, education of mom, number of siblings, permanent average local unemployment rate and wage rate at age 17.

(vii) interactions between the cohort dummies and corrected AFQT score, education of mom and number of siblings.

Outcome Equation

Following Carneiro et al. (2011), we make the assumption that the error terms are jointly independent of the covariates and the instruments. Then we have $\tau_{\texttt{MTE}}(a|\mathbf{x}) = \partial \mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}]/\partial a$, and

$$\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}] = \mathbf{x}^\mathsf{T}\boldsymbol{\gamma}_0 + a \cdot \mathbf{x}^\mathsf{T}\boldsymbol{\delta}_0 + \boldsymbol{\phi}(a)^\mathsf{T}\boldsymbol{\theta}_0,$$

where $P_i = \mathbb{P}[T_i = 1|\mathbf{Z}_i]$ is the propensity score, and $\boldsymbol{\phi}$ is some fixed transformation. To be more specific, we use series expansion of the estimated propensity score, with different order of polynomials (note that a linear term of the estimated propensity score is included in $a \cdot \mathbf{x}^\mathsf{T}$):

$p = 2$	$\boldsymbol{\phi}(a) = a^2$	Table 11
$p = 3$	$\boldsymbol{\phi}(a) = [a^2,\ a^3]^\mathsf{T}$	Table 12
$p = 4$	$\boldsymbol{\phi}(a) = [a^2,\ a^3,\ a^4]^\mathsf{T}$	Table 13
$p = 5$	$\boldsymbol{\phi}(a) = [a^2,\ a^3,\ a^4,\ a^5]^\mathsf{T}$	Table 14.

We use the same set of covariates $\mathbf{X}_i$ for the outcome equation as in Carneiro et al. (2011), which includes covariates (i)–(iv).

Selection Equation

The selection equation (i.e. the propensity score) is estimated with either a linear probability model or Logit model, and the dimension of $\mathbf{z}_i$ varies from 35 to 66. This is comparable to the simulation settings. We evaluate at the average values of the covariates, i.e. we report $\hat{\tau}_{\texttt{MTE}}(a|\bar{\mathbf{x}})$ with and without bias correction, for $a \in \{0.2, 0.5, 0.8\}$.

For selection equation, we consider five different specifications for $\mathbf{Z}_i$. The first one is most parsimonious, and corresponds to columns (1), (6) and (11) in Table 11–14, which includes (i)–(iii) and (v).

The next specification of $\mathbf{Z}_i$ include certain linear interactions, and corresponds to columns (2), (7) and (12) in Table 11–14, which includes (i)–(iii), (v) and (vi).

Another specification of $\mathbf{Z}_i$, corresponding to columns (3), (8) and (13) in Table 11–14, which includes (i)–(iii), (v) and (vii).

The next specification encompasses all above, given in columns (4), (9) and (14) in 11–14, which includes (i)–(iii) and (v)–(vii).

Finally, for comparison purpose, we also include the specification used in Carneiro et al. (2011), corresponding to columns (4), (10) and (15) in Table 11–14, where the propensity score is estimated with Logit regression employing (i)–(iii) and (v).

SA-9 Proofs

In this section we collect the technical proofs of lemmas, theorems and corollaries.

SA-9.1 Properties of $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^-\mathbf{Z}^\top$

Recall that $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^-\mathbf{Z}^\top$ is the projection matrix, with its entries denoted by π_{ij}. Then the first conclusion is that

$$\mathrm{tr}[\mathbf{\Pi}] = k.$$

And since $\mathbf{\Pi}$ is a projection matrix, one has $\mathbf{\Pi}\mathbf{\Pi} = \mathbf{\Pi}$, which means

$$\pi_{ij} = \sum_\ell \pi_{i\ell}\pi_{j\ell}.$$

Also, $\pi_{ij} = \pi_{ji}$ (i.e., $\mathbf{\Pi}$ is symmetric), and $0 \le \pi_{ii} \le 1$ from the idempotency of the projection matrix.

Next consider the trace of $\mathbf{\Pi}\mathbf{\Pi} = \mathbf{\Pi}^2$:

$$k = \mathrm{tr}[\mathbf{\Pi}^2] = \sum_i \sum_j \pi_{ij}^2 = \sum_i \pi_{ii}^2 + \sum_{i,j,j\neq i} \pi_{ij}^2,$$

which implies that

$$\sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j} \pi_{ij}^2 \le k.$$

Next we replace π_{ii} by $\sum_j \pi_{ij}^2$, which gives

$$k \ge \sum_i \pi_{ii}^2 = \sum_i \pi_{ii} \left(\sum_j \pi_{ij}^2 \right) = \sum_i \sum_j \pi_{ii}\pi_{ij}^2,$$

hence

$$\sum_i \pi_{ii}^3 \le k, \qquad \sum_{i,j} \pi_{ii}\pi_{ij}^2 \le k.$$

Now make a further replacement,

$$k \ge \sum_i \pi_{ii}^2 = \sum_i \left(\sum_j \pi_{ij}^2 \right)^2 = \sum_i \pi_{ii}^4 + \sum_{i,j,i\neq j} \pi_{ij}^4 + \sum_{i,j,\ell,j\neq\ell} \pi_{ij}^2\pi_{i\ell}^2.$$

One direct consequence is that

$$\sum_i \pi_{ii}^4 \le k, \qquad \sum_{i,j} \pi_{ij}^4 \le k, \qquad \sum_{i,j,\ell} \pi_{ij}^2\pi_{i\ell}^2 \le k.$$

We summarize the above in the following lemma:

Lemma SA.18.
Let $\mathbf{\Pi}$ be a projection matrix with rank at most k, then:
(i) $\mathbf{\Pi}$ is symmetric, nonnegative definite, and $\mathbf{\Pi}^2 = \mathbf{\Pi}$, which implies $\pi_{ij} = \sum_\ell \pi_{i\ell}\pi_{j\ell}$.
(ii) The diagonal elements satisfy

$$0 \le \pi_{ii} \le 1 \ \forall i, \qquad \text{and} \qquad \sum_i \pi_{ii} = \mathrm{tr}[\mathbf{\Pi}] \le k. \tag{E.34}$$

(iii) The following higher order summations hold:

$$\sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j} \pi_{ij}^2 \le k, \tag{E.35}$$

$$\sum_i \pi_{ii}^3 \le \sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j} \pi_{ii}\pi_{ij}^2 \le \sum_i \pi_{ii}^2 \le k, \tag{E.36}$$

$$\sum_i \pi_{ii}^4 \le \sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j} \pi_{ij}^4 \le \sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j,\ell} \pi_{ij}^2\pi_{i\ell}^2 \le \sum_i \pi_{ii}^2 \le k. \tag{E.37}$$

■

SA-9.2 Summation Expansion

We first consider the expansion of $(\sum_{i,j,i\neq j} a_{ij})^2$, where $a_{ij} \neq a_{ji}$.

$$\left(\sum_{i,j,i\neq j} a_{ij}\right)^2 = \sum_{\substack{i,j,i',j' \\ i\neq j,\ i'\neq j'}} a_{ij}a_{i'j'}$$

$$= \sum_{\substack{i,j,i',j' \\ \text{distinct}}} a_{ij}a_{i'j'} + \sum_{\substack{i,j,j' \\ \text{distinct}}} a_{ij}a_{ij'} + \sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{i'i} + \sum_{\substack{i,j,j' \\ \text{distinct}}} a_{ij}a_{jj'} + \sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{i'j} + \sum_{\substack{i,j \\ i\neq j}} a_{ij}^2 + \sum_{\substack{i,j \\ i\neq j}} a_{ij}a_{ji}.$$

Note that the two terms $\sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{i'i}$ and $\sum_{\substack{i,j,j' \\ \text{distinct}}} a_{ij}a_{jj'}$ are identical by relabeling, hence

Lemma SA.19.

$$\left(\sum_{i,j,i\neq j} a_{ij}\right)^2 = \sum_{\substack{i,j,i',j' \\ \text{distinct}}} a_{ij}a_{i'j'} + \sum_{\substack{i,j,j' \\ \text{distinct}}} a_{ij}a_{ij'} + 2\sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{i'i} + \sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{i'j} + \sum_{\substack{i,j \\ i\neq j}} a_{ij}^2 + \sum_{\substack{i,j \\ i\neq j}} a_{ij}a_{ji}. \tag{E.38}$$

■

A special case is when $a_{ij} = a_{ji}$ so that the two indices are exchangeable. Then

Lemma SA.20.

$$(i,j)\text{-exchangeable} \qquad \left(\sum_{i,j,i\neq j} a_{ij}\right)^2 = \sum_{\substack{i,j,i',j' \\ \text{distinct}}} a_{ij}a_{i'j'} + 4\sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{ii'} + 2\sum_{\substack{i,j \\ i\neq j}} a_{ij}^2. \tag{E.39}$$

■

Next we consider $(\sum_{\substack{i,j,\ell \\ \text{distinct}}} a_i b_{ij\ell})^2$, where $b_{ij\ell} = b_{i\ell j}$, i.e. for b the last two indices are exchangeable. For convenience define the following

$$d_i = \sum_{j,\ell,j\neq\ell} b_{ij\ell}, \qquad c_i = \sum_{\substack{j,\ell \\ j\neq i,\ell\neq i,j\neq\ell}} b_{ij\ell}.$$

Then

$$c_i = d_i - 2\sum_j b_{iij} + 2b_{iii} = d_i - 2\sum_{j,j\neq i} b_{iij}.$$

And the decomposition becomes

$$\left(\sum_{\substack{i,j,\ell \\ \text{distinct}}} a_i b_{ij\ell}\right)^2 = \left(\sum_i a_i c_i\right)^2 = \sum_i a_i^2 c_i^2 + \sum_{i,i',i\neq i'} a_i a_{i'} c_i c_{i'}.$$

To make further progress, consider

$$c_i^2 = \left(d_i - 2\sum_{j,j\neq i} b_{iij}\right)^2 = \left(\sum_{j,\ell,j\neq\ell} b_{ij\ell}\right)^2 + 4\left(\sum_{j,j\neq i} b_{iij}\right)^2 - 4\left(\sum_{j,\ell,j\neq\ell} b_{ij\ell}\right)\left(\sum_{\ell',\ell'\neq i} b_{ii\ell'}\right)$$
$$= \sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{ij'\ell'} + 4\sum_{\substack{j,\ell,j'\\ \text{distinct}}} b_{ij\ell}b_{ijj'} + 2\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}^2 + 4\sum_{j,j\neq i} b_{iij}^2 + 4\sum_{\substack{j,\ell\\ j\neq i,\ell\neq i,j\neq\ell}} b_{iij}b_{ii\ell} - 4\sum_{\substack{j,\ell,\ell'\\ j\neq\ell,\ell'\neq i}} b_{ij\ell}b_{ii\ell'},$$

and

$$c_ic_{i'} = \left(\sum_{j,\ell,j\neq\ell} b_{ij\ell}\right)\left(\sum_{j,\ell,j\neq\ell} b_{i'j\ell}\right) = \sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j'\ell'} + 4\sum_{\substack{j,\ell,\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j\ell'} + 2\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}b_{i'j\ell}.$$

Therefore we have the following

Lemma SA.21.

$$(j,\ell)\text{-exchangeable}\left(\sum_{\substack{i,j,\ell\\ \text{distinct}}} a_ib_{ij\ell}\right)^2$$
$$= \sum_i a_i^2\left[\sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{ij'\ell'} + 4\sum_{\substack{j,\ell,j'\\ \text{distinct}}} b_{ij\ell}b_{ijj'} + 2\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}^2\right] + 4\sum_i a_i^2\left[\sum_{j,j\neq i} b_{iij}^2 + \sum_{\substack{j,\ell\\ j\neq i,\ell\neq i,j\neq\ell}} b_{iij}b_{ii\ell}\right]$$
$$- 4\sum_i a_i^2\left[\sum_{\substack{j,\ell,\ell'\\ j\neq\ell,\ell'\neq i}} b_{ij\ell}b_{ii\ell'}\right] + \sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j'\ell'}\right] + 4\sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell,\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j\ell'}\right]$$
$$+ 2\sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}b_{i'j\ell}\right]. \tag{E.40}$$

■

SA-9.3 Theorem SA.1

Since $\hat{\boldsymbol{\theta}}$ is tight, let K be defined such that $\mathbb{P}\left[|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \geq K\right] \leq \eta$ for some $\eta > 0$. Then for an arbitrary $\delta > 0$

$$\mathbb{P}\left[|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \geq \delta\right] \leq \eta + \mathbb{P}\left[\delta \leq |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \leq K\right].$$

Define $G(\boldsymbol{\theta}) = G(\boldsymbol{\theta}, \mu) = |\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})]|$ and $G_n(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}, \hat{\mu}) = |n^{-1}\sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta})|$, then $\boldsymbol{\theta}_0 = \min_{\boldsymbol{\theta}} G(\boldsymbol{\theta})$, and $G_n(\hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}} G_n(\boldsymbol{\theta}) + o_{\mathbb{P}}(1)$. Further define $\varepsilon(\delta, K) = \inf_{\delta \leq |\boldsymbol{\theta}-\boldsymbol{\theta}_0| \leq K} G(\boldsymbol{\theta}) - G(\boldsymbol{\theta}_0)$, then $\varepsilon(\delta, K) > 0$ for all $\delta > 0$ and $K < \infty$, since we assumed $\boldsymbol{\theta}_0$ is the unique root and $\mathbf{m}$ is continuous in $\boldsymbol{\theta}$.

Note that $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \geq \delta$ and $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \leq K$ implies that either $|G(\boldsymbol{\theta}_0) - G_n(\boldsymbol{\theta}_0)| \geq \varepsilon(\delta, K)/3 + o_{\mathbb{P}}(1)$, or $|G(\hat{\boldsymbol{\theta}}) - G_n(\hat{\boldsymbol{\theta}})| \geq \varepsilon(\delta, K)/3 + o_{\mathbb{P}}(1)$. Therefore

$$\mathbb{P}\left[\delta \leq |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \leq K\right] \leq \mathbb{P}\left[\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0| \leq K} |G_n(\boldsymbol{\theta}) - G(\boldsymbol{\theta})| + o_{\mathbb{P}}(1) \geq \varepsilon(\delta, K)/3\right]$$
$$\leq \mathbb{1}\left[\sup_{\substack{|\boldsymbol{\theta}-\boldsymbol{\theta}_0| \leq K\\ \max_{1\leq i\leq n}|\mu_i' - \mu_i| \leq \lambda}} |G(\boldsymbol{\theta}, \mu') - G(\boldsymbol{\theta})| \geq \varepsilon(\delta, K)/6\right]$$

$$+\mathbb{P}\left[\max_{1\le i\le n}|\hat{\mu}_i-\mu_i|\ge\lambda\right]+\mathbb{P}\left[\sup_{\substack{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le K\\ \max_{1\le i\le n}|\mu_i'-\mu_i|\le\lambda}}\left|G_n(\boldsymbol{\theta},\mu')-G(\boldsymbol{\theta},\mu')\right|+o_{\mathbb{P}}(1)\le\varepsilon(\delta,K)/6\right].$$

By Assumption A.2(1), one has $\limsup_n\mathbb{P}\left[\max_{1\le i\le n}|\hat{\mu}_i-\mu_i|\ge\lambda\right]=0$ for any (fixed) $\lambda>0$. Further, due to Assumption A.1(4),

$$\limsup_n\mathbb{P}\left[\sup_{\substack{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le K\\ \max_{1\le i\le n}|\mu_i'-\mu_i|\le\lambda}}\left|G_n(\boldsymbol{\theta},\mu')-G(\boldsymbol{\theta},\mu')\right|+o_{\mathbb{P}}(1)\le\varepsilon(\delta,K)/6\right]=0.$$

Therefore

$$\limsup_n\mathbb{P}\left[|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0|\ge\delta\right]\le\eta+\limsup_n\mathbb{1}\left[\sup_{\substack{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le K\\ \max_{1\le i\le n}|\mu_i'-\mu_i|\le\lambda}}|G(\boldsymbol{\theta},\mu')-G(\boldsymbol{\theta})|\ge\varepsilon(\delta,K)/6\right].$$

Finally, note that K implicitly depends on η, while the choice of δ, η and λ are mutually independent. Hence we could first let $\lambda\downarrow 0$, then the indicator function will be identically zero for all n (use the dominated convergence theorem). Then let $\eta\downarrow 0$, we will have the desired consistency result. ■

SA-9.4 Lemma SA.2

We apply Taylor expansion to the GMM problem, which gives

$$\begin{aligned}o_{\mathbb{P}}(1)&=\left[\frac{1}{n}\sum_i\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})\right]^{\mathsf{T}}\boldsymbol{\Omega}_n\frac{1}{\sqrt{n}}\sum_i\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})\\&=\left[\frac{1}{n}\sum_i\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})\right]^{\mathsf{T}}\boldsymbol{\Omega}_n\left(\frac{1}{\sqrt{n}}\sum_i\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)+\left[\frac{1}{n}\sum_i\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\right]\sqrt{n}\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big)\right),\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ is (possibly random) convex combination of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Then we have

$$\begin{aligned}\sqrt{n}\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big)&=-(\hat{\mathbf{M}}_n^{\mathsf{T}}\boldsymbol{\Omega}_n\tilde{\mathbf{M}}_n)^{-1}\hat{\mathbf{M}}_n^{\mathsf{T}}\boldsymbol{\Omega}_n\frac{1}{\sqrt{n}}\sum_i\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)+o_{\mathbb{P}}(1)\\&=-(\mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_0\mathbf{M}_0)^{-1}\mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_0\frac{1}{\sqrt{n}}\sum_i\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)+o_{\mathbb{P}}(1),\end{aligned}$$

where

$$\hat{\mathbf{M}}_n=\frac{1}{n}\sum_i\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}),\qquad\tilde{\mathbf{M}}_n=\frac{1}{n}\sum_i\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}}).$$

In the above, we used the fact that both $\hat{\mathbf{M}}_n$ and $\tilde{\mathbf{M}}_n$ converge in probability to $\mathbf{M}_0$. This is easily shown by noting that (c.f. Assumption A.1(5))

$$\left|\hat{\mathbf{M}}_n-\frac{1}{n}\sum_i\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\right|\le\Big(\frac{1}{n}\sum_i\mathcal{H}_i^{\alpha,\delta}(\partial\mathbf{m}/\partial\boldsymbol{\theta})\Big)\cdot\Big(\max_{1\le i\le n}|\hat{\mu}_i-\mu_i|+|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0|\Big)^{\alpha}=o_{\mathbb{P}}(1),$$

since $\hat{\mu}_i$ is uniformly consistent and $\hat{\boldsymbol{\theta}}$ is consistent. And note that $n^{-1}\sum_i\partial\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^{\mathsf{T}}\to_{\mathbb{P}}\mathbf{M}_0$ by the law of large numbers. The same argument applies to $\tilde{\mathbf{M}}_n$. ■

SA-9.5 Lemma SA.3

SA-9.5.1 Approximation Bias

For simplicity, let $\dot{\mathbf{m}}_i = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$, then

$$\left|\frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}_i\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right| \le \left|\frac{1}{\sqrt{n}}\sum_i \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right| + \left|\frac{1}{\sqrt{n}}\sum_i (\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i])\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right|,$$

and we call the two terms (I) and (II) respectively. For term (I), we use projection matrix property, which implies

$$\text{(I)} = \left|\frac{1}{\sqrt{n}}\sum_i \eta_i\Big(\mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i] - \sum_j \pi_{ij}\mathbb{E}[\dot{\mathbf{m}}_j|\mathbf{z}_j]\Big)\right| \le \sqrt{n}\sqrt{\frac{1}{n}\sum_i \eta_i^2}\sqrt{\frac{1}{n}\sum_i \Big|\mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i] - \sum_j \pi_{ij}\mathbb{E}[\dot{\mathbf{m}}_j|\mathbf{z}_j]\Big|^2}.$$

By further splitting the conditional expectation $\mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]$ into a linear projection and an error term,

$$\begin{aligned}\text{(I)} &\le \sqrt{n}\sqrt{\frac{1}{n}\sum_i \eta_i^2}\sqrt{\frac{1}{n}\sum_i \Big|\boldsymbol{\zeta}_i - \sum_j \pi_{ij}\boldsymbol{\zeta}_j\Big|^2} \le \sqrt{n}\sqrt{\frac{1}{n}\sum_i \eta_i^2}\sqrt{\frac{1}{n}\sum_i \Big|\boldsymbol{\zeta}_i\Big|^2} \\ &= O_{\mathbb{P}}\left(\sqrt{n\mathbb{E}[\eta_i^2]\mathbb{E}[|\boldsymbol{\zeta}_i|^2]}\right) = o_{\mathbb{P}}(1).\end{aligned}$$

The second term (II) can be bounded with conditional expectation and variance calculations. First note that since η_i is the error from linear approximation, this term has zero conditional mean:

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_i (\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i])\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right] = \frac{1}{\sqrt{n}}\sum_i \mathbb{E}[\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]|\mathbf{z}_i]\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big) = \mathbf{0}.$$

Next we consider the conditional second moment:

$$\begin{aligned}&\left|\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_i (\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i])\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right]\right| \asymp \frac{1}{n}\sum_i \Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)^2 \mathbb{E}[|\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]|^2|\mathbf{z}_i] \\ &\asymp \frac{1}{n}\sum_i \Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)^2 \le \frac{1}{n}\sum_i \eta_i^2 = O_{\mathbb{P}}\left(\mathbb{E}[\eta_i^2]\right) = o_{\mathbb{P}}(1),\end{aligned}$$

where for the second line, we used the assumption that $\dot{\mathbf{m}}_i$ has uniformly bounded conditional variance. ■

SA-9.5.2 Influence Function and Asymptotic Bias

The conclusion will be self-evident after two decompositions. First rewrite $\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) - \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i\right] + \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i\right]$ as the conditional expectation decomposition. Then

$$\text{(E.9)} = \frac{1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right)\cdot\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij},$$

where we use $\mathbf{u}_i = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) - \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i\right]$ to save notation. Then

$$\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij} = \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right] + O_{\mathbb{P}}\left(\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right]^{1/2}\right),$$

where we use $\mathbb{E}_{[\cdot|\mathbf{Z}]}$ and $\mathbb{V}_{[\cdot|\mathbf{Z}]}$ to denote the expectation and variance conditional on $\{\mathbf{z}_i, \mu_i\}_{1\le i\le n}$, respectively. Then

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right] = \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{1,i}\pi_{ii},$$

with $\mathbf{b}_{1,i} = \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{u}_i\varepsilon_i] = \mathbb{E}_{[\cdot|\mathbf{Z}]}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\varepsilon_i]$, since

$$i \ne j \qquad \Rightarrow \qquad \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\varepsilon_j\right] = \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{u}_i]\cdot\mathbb{E}_{[\cdot|\mathbf{Z}]}[\varepsilon_j] = \mathbf{0}.$$

Next we estimate the order of the conditional variance. To this end, consider

$$
\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)^{\mathsf{T}}\right] \\
&= \frac{1}{n}\sum_{i,j,i',j'}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i'}^{\mathsf{T}}\varepsilon_j\varepsilon_{j'}\pi_{ij}\pi_{i'j'}\right] \\
&= \frac{1}{n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i'}^{\mathsf{T}}\varepsilon_i\varepsilon_{i'}\pi_{ii}\pi_{i'i'}\right] && (i=j,\ i'=j') \\
&+ \frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathsf{T}}\varepsilon_j\varepsilon_{j}\pi_{ij}\pi_{ij}\right] && (i=i',\ j=j') \\
&+ \frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{j}^{\mathsf{T}}\varepsilon_j\varepsilon_{i}\pi_{ij}\pi_{ij}\right] && (i=j',\ j=i') \\
&+ \frac{1}{n}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathsf{T}}\varepsilon_i\varepsilon_{i}\pi_{ii}\pi_{ii}\right]. && (i=j=i'=j')
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right] \\
&= \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)^{\mathsf{T}}\right] - \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right]\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right]^{\mathsf{T}} \\
&= \frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathsf{T}}\varepsilon_j\varepsilon_{j}\pi_{ij}\pi_{ij}\right] + \frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{j}^{\mathsf{T}}\varepsilon_j\varepsilon_{i}\pi_{ij}\pi_{ij}\right] + \frac{1}{n}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathsf{T}}\varepsilon_i\varepsilon_{i}\pi_{ii}\pi_{ii}\right] - \frac{1}{n}\sum_{i}\mathbf{b}_{1,i}\mathbf{b}_{1,i}^{\mathsf{T}}\pi_{ii}^2.
\end{aligned}
$$

Due to Assumption A.1(7), the above terms are easily bounded by

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathsf{T}}\varepsilon_j\varepsilon_{j}\pi_{ij}\pi_{ij}\right]\right| \precsim \frac{1}{n}\sum_{i,j}\pi_{ij}^2 \leq \frac{k}{n} \\
&\left|\frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{j}^{\mathsf{T}}\varepsilon_j\varepsilon_{i}\pi_{ij}\pi_{ij}\right]\right| \precsim \frac{1}{n}\sum_{i,j}\pi_{ij}^2 \leq \frac{k}{n} \\
&\left|\frac{1}{n}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathsf{T}}\varepsilon_i\varepsilon_{i}\pi_{ii}\pi_{ii}\right]\right| \precsim \frac{1}{n}\sum_{i}\pi_{ii}^2 \leq \frac{k}{n} \\
&\left|\frac{1}{n}\sum_{i}\mathbf{b}_{1,i}\mathbf{b}_{1,i}^{\mathsf{T}}\pi_{ii}^2\right| \precsim \frac{1}{n}\sum_{i}\pi_{ii}^2 \leq \frac{k}{n},
\end{aligned}
$$

which closes the proof. ■

SA-9.5.3 Variance Simplification

For notational convenience, denote $\mathbf{a}_i = \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]$. Then it suffices to give conditions such that

$$
\frac{1}{\sqrt{n}}\sum_{i}\left[\mathbf{a}_i - \sum_{j}\mathbf{a}_j\pi_{ij}\right]\varepsilon_i = o_{\mathbb{P}}(1).
$$

Note that the conditional variance of the LHS is (use Assumption A.1(7))

$$\begin{aligned}
&\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\left|\frac{1}{\sqrt{n}}\sum_i\left[\mathbf{a}_i-\sum_j\mathbf{a}_j\pi_{ij}\right]\varepsilon_i\right|\right] \precsim \frac{1}{n}\sum_i\left|\mathbf{a}_i-\sum_j\mathbf{a}_j\pi_{ij}\right|^2 = \frac{1}{n}\sum_i\left|\mathbf{a}_i-\mathbf{\Gamma}\mathbf{z}_i+\mathbf{\Gamma}\mathbf{z}_i-\sum_j\mathbf{a}_j\pi_{ij}\right|^2 \\
&\leq \frac{2}{n}\sum_i\left(\left|\mathbf{a}_i-\mathbf{\Gamma}\mathbf{z}_i\right|^2+\left|\mathbf{\Gamma}\mathbf{z}_i-\sum_j\mathbf{a}_j\pi_{ij}\right|^2\right) = \frac{2}{n}\sum_i\left|\mathbf{a}_i-\mathbf{\Gamma}\mathbf{z}_i\right|^2+\frac{2}{n}\sum_i\left|\sum_j(\mathbf{a}_j-\mathbf{\Gamma}\mathbf{z}_j)\pi_{ij}\right|^2 \\
&\leq \frac{4}{n}\sum_i\left|\mathbf{a}_i-\mathbf{\Gamma}\mathbf{z}_i\right|^2 \qquad \text{(Projection)} \\
&= o_{\mathbb{P}}(1),
\end{aligned}$$

where the last line shows why the assumption in Lemma SA.3 is sufficient. Note that by projection, $\mathbf{\Gamma}\mathbf{z}_i = \sum_j \mathbf{\Gamma}\mathbf{z}_j\pi_{ij}$. ■

SA-9.6 Lemma SA.4

SA-9.6.1 Approximation Error

For the current proof, we use $\ddot{\mathbf{m}}_i = \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$ for notational convenience. Then recall that $\hat{\mu}_i - \mu_i = \sum_j \pi_{ij}\varepsilon_j - (\eta_i - \sum_j \pi_{ij}\eta_j)$. Then

$$\begin{aligned}
\text{(E.10)} &= \frac{1}{2\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\left(\sum_j\pi_{ij}\varepsilon_j-(\eta_i-\sum_j\pi_{ij}\eta_j)\right)^2 \\
&= \underbrace{\frac{1}{2\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\left(\sum_j\pi_{ij}\varepsilon_j\right)^2}_{\text{(I)}} + \underbrace{\frac{1}{2\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2}_{\text{(II)}} - \underbrace{\frac{1}{\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\left(\sum_j\pi_{ij}\varepsilon_j\right)\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)}_{\text{(III)}}.
\end{aligned}$$

We first deal with (II). Again we make a conditional expectation expansion of $\ddot{\mathbf{m}}_i$, which implies

$$|\text{(II)}| \leq \underbrace{\left|\frac{1}{2\sqrt{n}}\sum_i \mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right|}_{\text{(II.1)}} + \underbrace{\left|\frac{1}{2\sqrt{n}}\sum_i (\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i])\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right|}_{\text{(II.2)}}.$$

(II.1) has the simple bound:

$$\text{(II.1)} \leq \frac{1}{2\sqrt{n}}\sum_i |\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]|\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2 \precsim \frac{1}{\sqrt{n}}\sum_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2 \leq \frac{1}{\sqrt{n}}\sum_i\eta_i^2 = O_{\mathbb{P}}(\sqrt{n}\mathbb{E}[\eta_i^2]) = o_{\mathbb{P}}(1),$$

where we used the assumption that $\ddot{\mathbf{m}}_i$ has uniformly bounded conditional expectation.

For (II.2), we employ conditional expectation and variance calculation. Note that it has zero conditional expectation:

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{2\sqrt{n}}\sum_i(\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i])\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right] = \frac{1}{2\sqrt{n}}\sum_i \mathbb{E}[\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]|\mathbf{z}_i]\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2 = \mathbf{0}.$$

The conditional variance is bounded by the following:

$$\begin{aligned}
&\left|\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{2\sqrt{n}}\sum_i(\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i])\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right]\right| \precsim \frac{1}{n}\sum_i \mathbb{E}[|\ddot{\mathbf{m}}_i|^2|\mathbf{z}_i]\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^4 \\
&\precsim \frac{1}{n}\sum_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^4 = \frac{1}{n}\sum_i\check{\eta}_i^4,
\end{aligned}$$

where in the second line we used the assumption that $\ddot{\mathbf{m}}_i$ has uniformly bounded conditional second moment, and

we use $\check{\eta}_i = \eta_i - \sum_j \pi_{ij}\eta_j$ for simplicity. Next, note that

$$\frac{1}{n}\left(\sum_i \check{\eta}_i^4 + \sum_{i,j,i\neq j} \check{\eta}_i^2\check{\eta}_j^2\right) = \left(\frac{1}{\sqrt{n}}\sum_i \check{\eta}_i^2\right)^2 \leq \left(\frac{1}{\sqrt{n}}\sum_i \eta_i^2\right)^2 = o_{\mathbb{P}}(1),$$

so that we conclude the previous conditional variance is asymptotically negligible.

For term (III), we first compute its conditional expectation:

$$\begin{aligned}
&\left|\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\left(\sum_j \pi_{ij}\varepsilon_j\right)\left(\eta_i - \sum_j \pi_{ij}\eta_j\right)\right]\right| = \left|\frac{1}{\sqrt{n}}\sum_i \pi_{ii}\mathbb{E}[\ddot{\mathbf{m}}_i\varepsilon_i|\mathbf{z}_i]\left(\eta_i - \sum_j \pi_{ij}\eta_j\right)\right| \\
&\leq \sqrt{n}\sqrt{\frac{1}{n}\sum_i \pi_{ii}^2|\mathbb{E}[\ddot{\mathbf{m}}_i\varepsilon_i|\mathbf{z}_i]|^2}\sqrt{\frac{1}{n}\sum_i\left(\eta_i - \sum_j \pi_{ij}\eta_j\right)^2} \precsim \sqrt{n}\sqrt{\frac{1}{n}\sum_i \pi_{ii}^2}\sqrt{\frac{1}{n}\sum_i\left(\eta_i - \sum_j \pi_{ij}\eta_j\right)^2} \\
&= o_{\mathbb{P}}\left(\sqrt{n}\sqrt{\frac{k}{n}}\frac{1}{n^{1/4}}\right) = o_{\mathbb{P}}\left(\sqrt{\frac{k}{\sqrt{n}}}\right) = o_{\mathbb{P}}(1).
\end{aligned}$$

Here for the second line, we use the assumption that $\mathbb{E}[\ddot{\mathbf{m}}_i\varepsilon_i|\mathbf{z}_i]$ is uniformly bounded. Hence to bound (III), it suffices to consider the conditional second moment, which is bounded by the following (where $\check{\eta}_i = \eta_i - \sum_j \pi_{ij}\eta_j$):

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{n}\sum_{i,j,k,\ell}|\ddot{\mathbf{m}}_i||\ddot{\mathbf{m}}_j|\varepsilon_k\varepsilon_\ell\pi_{ik}\pi_{j\ell}\check{\eta}_i\check{\eta}_j\right] \\
&= \frac{1}{n}\sum_{\substack{i,j,\ell \\ \text{distinct}}} \mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\mathbb{E}[|\ddot{\mathbf{m}}_j||\mathbf{z}_j]\mathbb{E}[\varepsilon_\ell^2|\mathbf{z}_\ell]\pi_{i\ell}\pi_{j\ell}\check{\eta}_i\check{\eta}_j && \text{(III.1: } k=\ell\text{)} \\
&+ \frac{1}{n}\sum_{\substack{i,j \\ \text{distinct}}} \mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\mathbb{E}[|\ddot{\mathbf{m}}_j|\varepsilon_j^2|\mathbf{z}_j]\pi_{ij}\pi_{jj}\check{\eta}_i\check{\eta}_j && \text{(III.2: } j=k=\ell\text{)} \\
&+ \frac{1}{n}\sum_{\substack{i,\ell \\ \text{distinct}}} \mathbb{E}[|\ddot{\mathbf{m}}_i|^2|\mathbf{z}_i]\mathbb{E}[\varepsilon_\ell^2|\mathbf{z}_\ell]\pi_{i\ell}^2\check{\eta}_i^2 && \text{(III.3: } i=j,k=\ell\text{)} \\
&+ o_{\mathbb{P}}(1), && (i=k, j=\ell)
\end{aligned}$$

where the last $o_{\mathbb{P}}(1)$ is the squared conditional expectation, and has been handled earlier. (III.3) is the simplest, which has bound

$$\text{(III.3)} \precsim \frac{1}{n}\sum_{i,\ell}\pi_{i\ell}^2\check{\eta}_i^2 = \frac{1}{n}\sum_i \pi_{ii}\check{\eta}_i^2 \leq \frac{1}{n}\sum_i \check{\eta}_i^2 \leq \frac{1}{n}\sum_i \eta_i^2 = o_{\mathbb{P}}(1).$$

(III.1) is also easy, since by projection property, one has (it is easier to write it into a quadratic matrix form)

$$\text{(III.1)} \precsim \frac{1}{n}\sum_{i,j,\ell}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\mathbb{E}[|\ddot{\mathbf{m}}_j||\mathbf{z}_j]\pi_{i\ell}\pi_{j\ell}\check{\eta}_i\check{\eta}_j \leq \frac{1}{n}\sum_i |\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]|^2|\check{\eta}|_i^2 \precsim \frac{1}{n}\sum_i \check{\eta}_i^2 \leq \frac{1}{n}\sum_i \eta_i^2 = o_{\mathbb{P}}(1).$$

(III.2) is bounded by the following:

$$\begin{aligned}
\text{(III.2)} &\precsim \left|\frac{1}{n}\sum_j \mathbb{E}[|\ddot{\mathbf{m}}_j|\varepsilon_j^2|\mathbf{z}_j]\pi_{jj}\check{\eta}_j\sum_i \mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\pi_{ij}\check{\eta}_i\right| \\
&\leq \sqrt{\frac{1}{n}\sum_j(\mathbb{E}[|\ddot{\mathbf{m}}_j|\varepsilon_j^2|\mathbf{z}_j])^2\pi_{jj}^2\check{\eta}_j^2}\sqrt{\frac{1}{n}\sum_j\left(\sum_i \mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\pi_{ij}\check{\eta}_i\right)^2} \\
&\precsim \sqrt{\frac{1}{n}\sum_j \eta_j^2}\sqrt{\frac{1}{n}\sum_{j,i,\ell}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\pi_{ij}\check{\eta}_i\mathbb{E}[|\ddot{\mathbf{m}}_\ell||\mathbf{z}_\ell]\pi_{\ell j}\check{\eta}_\ell} \\
&= \sqrt{\frac{1}{n}\sum_j \eta_j^2}\sqrt{\frac{1}{n}\sum_{i,\ell}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\check{\eta}_i\pi_{i\ell}\mathbb{E}[|\ddot{\mathbf{m}}_\ell||\mathbf{z}_\ell]\check{\eta}_\ell}
\end{aligned}$$

$$\leq \sqrt{\frac{1}{n}\sum_j \eta_j^2}\sqrt{\frac{1}{n}\sum_i (\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\check{\eta}_i)^2} \precsim \sqrt{\frac{1}{n}\sum_j \eta_j^2}\sqrt{\frac{1}{n}\sum_i \check{\eta}_i^2}$$
$$\leq \frac{1}{n}\sum_j \eta_j^2 = o_{\mathbb{P}}(1),$$

which concludes the proof. ■

SA-9.6.2 Asymptotic Bias

Again we define $\ddot{\mathbf{m}}_i = \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$ to save notation. For the proof again we consider the expansion

$$\frac{1}{2\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\left(\sum_j \pi_{ij}\varepsilon_j\right)^2 = \frac{1}{2\sqrt{n}}\sum_{i,j,\ell}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell$$
$$= \underbrace{\frac{1}{2\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell}_{\text{(I)}} + \underbrace{\frac{1}{2\sqrt{n}}\sum_{i,j,i\neq j}\ddot{\mathbf{m}}_i\pi_{ij}^2\varepsilon_j^2}_{(II)} + \underbrace{\frac{2}{2\sqrt{n}}\sum_{i,j,i\neq j}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{ii}\varepsilon_i\varepsilon_j}_{(III)} + \underbrace{\frac{1}{2\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2}_{(Iv)}.$$

Expectation

It is easy to see that both (I) and (III) have zero conditional expectation. Hence we consider (II) and (IV).

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\text{II})\right] = \frac{1}{2\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\pi_{ij}^2\varepsilon_j^2\right] = \frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbf{b}_{2,ij}\pi_{ij}^2.$$

where the last line uses (E.35). And

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\text{IV})\right] = \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{2,ii}\pi_{ii}^2.$$

Variance, Term (I)

First for (I) we use (E.40) with $a_i = \ddot{\mathbf{m}}_i$ and (ignore the 1/2 in front) $b_{ij\ell} = \pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell$, and

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell\right)^2\right]$$
$$= \underbrace{\frac{2}{n}\sum_i\sum_{j,\ell,j\neq\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_i b_{ij\ell}^2\right]}_{\text{(I.1)}} + \underbrace{\frac{4}{n}\sum_i\sum_{j,j\neq i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_i b_{iij}^2\right]}_{\text{(I.2)}} + \underbrace{\frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j,\ell,j\neq\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_{i'} b_{ij\ell}b_{i'jl}\right]}_{\text{(I.3)}}.$$

Next by (E.37) and (E.36), respectively,

$$\text{(I.1)} \precsim \frac{1}{n}\sum_i\sum_{j,\ell,j\neq\ell}\pi_{ij}^2\pi_{i\ell}^2 \leq \frac{1}{n}\sum_i \pi_{ii}^2 \leq \frac{k}{n}.$$
$$\text{(I.2)} \precsim \frac{1}{n}\sum_{i,j}\pi_{ii}^2\pi_{ij}^2 \leq \frac{1}{n}\sum_i \pi_{ii}^3 \leq \frac{k}{n}.$$

And

$$\text{(I.3)} = \frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j,j',j\neq j'}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_{i'}\pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\varepsilon_j^2\varepsilon_{j'}^2\right] = \frac{2}{n}\sum_{\substack{i,i',j,j'\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_{i'}\pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\varepsilon_j^2\varepsilon_{j'}^2\right]$$
$$+ \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_{i'}\pi_{ii}\pi_{ii'}^2\pi_{i'i'}\varepsilon_i^2\varepsilon_{i'}^2\right] + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_{i'}\pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\varepsilon_i^2\varepsilon_j^2\right].$$

Define $\mathbf{c}_i = \mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]$, $d_j = \mathbb{E}[\varepsilon_j^2|\mathbf{z}_j]$, and $\mathbf{e}_i = \mathbb{E}[\varepsilon_i^2\ddot{\mathbf{m}}_i|\mathbf{z}_i]$, and with (E.40) the above becomes

$$\begin{aligned}
\text{(I.3)} &= \frac{2}{n}\sum_{\substack{i,i',j,j'\\ \text{distinct}}} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}d_jd_{j'} + \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\mathbf{e}_i^\mathsf{T}\mathbf{e}_{i'} + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\mathbf{e}_i^\mathsf{T}\mathbf{c}_{i'}d_j \\
&= \frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j,j',j\neq j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}d_jd_{j'} \\
&+ \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^\mathsf{T}\mathbf{e}_{i'} - \mathbf{c}_i^\mathsf{T}d_i\mathbf{c}_{i'}d_{i'}\right) + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_id_i\right)^\mathsf{T}\mathbf{c}_{i'}d_j \\
&= \underbrace{\frac{2}{n}\sum_{i,i',j,j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}d_jd_{j'}}_{\text{(I.3.1)}} + \underbrace{\frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^\mathsf{T}\mathbf{e}_{i'} - \mathbf{c}_i^\mathsf{T}d_i\mathbf{c}_{i'}d_{i'}\right)}_{\text{(I.3.2)}} \\
&+ \underbrace{\frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_id_i\right)^\mathsf{T}\mathbf{c}_{i'}d_j}_{\text{I.3.3}} - \underbrace{\frac{2}{n}\sum_{i,i',i\neq i'}\sum_j \pi_{ij}^2\pi_{i'j}^2\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}d_j^2}_{\text{(I.3.4)}} - \underbrace{\frac{2}{n}\sum_i\sum_{j,j'}\pi_{ij}^2\pi_{ij'}^2|\mathbf{c}_i|^2d_jd_{j'}}_{\text{(I.3.5)}}.
\end{aligned}$$

Then use (E.35)

$$\begin{aligned}
|\text{(I.3.1)}| &= \left|\frac{2}{n}\sum_{i,i',j,j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}d_jd_{j'}\right| = \left|\frac{2}{n}\sum_{i,i'}\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}\left(\sum_j \pi_{ij}\pi_{i'j}d_j\right)^2\right| \leq \max_{1\leq i,i'\leq n}|\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}|\frac{2}{n}\sum_{i,i'}\left(\sum_j \pi_{ij}\pi_{i'j}d_j\right)^2 \\
&= \max_{1\leq i,i'\leq n}|\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}|\frac{2}{n}\sum_{i,i',j,j'}\pi_{ij}\pi_{ij'}d_jd_{j'} = \max_{1\leq i,i'\leq n}|\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}|\frac{2}{n}\sum_{j,j'}d_jd_{j'}\left(\sum_i \pi_{ij}\pi_{ij'}\right)^2 \\
&\leq \max_{1\leq i,i',j,j'\leq n}|\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}d_jd_{j'}|\frac{2}{n}\sum_{j,j'}\left(\sum_i \pi_{ij}\pi_{ij'}\right)^2 \leq \max_{1\leq i,i',j,j'\leq n}|\mathbf{c}_i^\mathsf{T}\mathbf{c}_{i'}d_jd_{j'}|\frac{2}{n}\sum_{j,j'}\pi_{jj'}^2 \asymp \frac{k}{n}.
\end{aligned}$$

And by (E.36)

$$\begin{aligned}
|\text{(I.3.2)}| &= \left|\frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^\mathsf{T}\mathbf{e}_{i'} - \mathbf{c}_i^\mathsf{T}d_i\mathbf{c}_{i'}d_{i'}\right)\right| \leq \max_{1\leq i,i'\leq n}|\mathbf{e}_i^\mathsf{T}\mathbf{e}_{i'} - \mathbf{c}_i^\mathsf{T}d_i\mathbf{c}_{i'}d_{i'}|\frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'} \\
&\leq \max_{1\leq i,i'\leq n}|\mathbf{e}_i^\mathsf{T}\mathbf{e}_{i'} - \mathbf{c}_i^\mathsf{T}d_i\mathbf{c}_{i'}d_{i'}|\frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii'}^2\pi_{i'i'} \asymp \frac{k}{n}.
\end{aligned}$$

And by (E.35) and (E.37)

$$\begin{aligned}
|\text{(I.3.3)}| &= \left|\frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_id_i\right)^\mathsf{T}\mathbf{c}_{i'}d_j\right| \asymp \frac{1}{n}\sum_{i',j,i'\neq j}|\mathbf{c}_{i'}d_j||\pi_{i'j}|\left|\sum_{\substack{i\\ i\neq i',i\neq j}}(\mathbf{e}_i - \mathbf{c}_id_i)\pi_{ii}\pi_{ij}\pi_{ii'}\right| \\
&\asymp \frac{1}{n}\sqrt{\sum_{i',j}\pi_{i'j}^2}\sqrt{\sum_{i',j}\left|\sum_{\substack{i\\ i\neq i',i\neq j}}(\mathbf{e}_i - \mathbf{c}_id_i)\pi_{ii}\pi_{ij}\pi_{ii'}\right|^2} \asymp \frac{\sqrt{k}}{n}\sqrt{\sum_{i,i',j,j'}(\mathbf{e}_i - \mathbf{c}_id_i)^\mathsf{T}(\mathbf{e}_{j'} - \mathbf{c}_{j'}d_{j'})\pi_{ii}\pi_{ij}\pi_{ii'}\pi_{j'j'}\pi_{jj'}\pi_{i'j'}} \\
&= \frac{\sqrt{k}}{n}\sqrt{\sum_{i,j'}(\mathbf{e}_i - \mathbf{c}_id_i)^\mathsf{T}(\mathbf{e}_{j'} - \mathbf{c}_{j'}d_{j'})\pi_{ii}\pi_{ij'}^2\pi_{j'j'}} \asymp \frac{\sqrt{k}}{n}\sqrt{\sum_{i,j'}\pi_{ii}\pi_{ij'}^2\pi_{j'j'}} \leq \frac{\sqrt{k}}{n}\sqrt{\sum_{i,j'}\pi_{ii}\pi_{ij'}^2} \leq \frac{\sqrt{k}}{n}\sqrt{\sum_i \pi_{ii}^2} \leq \frac{k}{n}.
\end{aligned}$$

And by (E.37)

$$|(\text{I.3.4})| = \left| \frac{2}{n} \sum_{i,i',i\neq i'} \sum_{j} \pi_{ij}^2 \pi_{i'j}^2 \mathbf{c}_i^\top \mathbf{c}_{i'} d_j^2 \right| \leq \max_{1\leq i,i',j\leq n} |\mathbf{c}_i^\top \mathbf{c}_{i'} d_j^2| \frac{2}{n} \sum_{i,i',i\neq i'} \sum_{j} \pi_{ij}^2 \pi_{i'j}^2 \asymp \frac{k}{n}.$$

And by (E.37)

$$|(\text{I.3.5})| = \left| \frac{2}{n} \sum_{i} \sum_{j,j'} \pi_{ij}^2 \pi_{ij'}^2 |\mathbf{c}_i|^2 d_j d_{j'} \right| \leq \max_{1\leq i,j,j'\leq n} ||\mathbf{c}_i|^2 d_j d_{j'}| \frac{2}{n} \sum_{i} \sum_{j,j'} \pi_{ij}^2 \pi_{ij'}^2 \asymp \frac{k}{n}.$$

Variance, Term (II)

Then for (II), one has (by using (E.38))

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\left(\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \ddot{\mathbf{m}}_i \pi_{ij}^2 \varepsilon_j^2 \right)^2 \right] - \left(\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \pi_{ij}^2 \varepsilon_j^2 \right] \right)^2 \\
&= \underbrace{\frac{1}{n} \sum_{\substack{i,i',j,j' \\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_{i'} \pi_{ij}^2 \pi_{i'j'}^2 \varepsilon_j^2 \varepsilon_{j'}^2 \right] - \left(\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \pi_{ij}^2 \varepsilon_j^2 \right] \right)^2}_{\text{(II.1)}} \\
&+ \underbrace{\frac{1}{n} \sum_{\substack{i,j,j' \\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[|\ddot{\mathbf{m}}_i|^2 \pi_{ij}^2 \pi_{ij'}^2 \varepsilon_j^2 \varepsilon_{j'}^2 \right]}_{\text{(II.2)}} + \underbrace{\frac{2}{n} \sum_{\substack{i,i',j \\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_{i'} \pi_{ij}^2 \pi_{ii'}^2 \varepsilon_i^2 \varepsilon_j^2 \right]}_{\text{(II.3)}} + \underbrace{\frac{1}{n} \sum_{\substack{i,i',j \\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_{i'} \pi_{ij}^2 \pi_{i'j}^2 \varepsilon_j^4 \right]}_{\text{(II.4)}} \\
&+ \underbrace{\frac{1}{n} \sum_{i,j,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[|\ddot{\mathbf{m}}_i|^2 \pi_{ij}^4 \varepsilon_j^4 \right]}_{\text{(II.5)}} + \underbrace{\frac{1}{n} \sum_{i,j,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_j \pi_{ij}^4 \varepsilon_i^2 \varepsilon_j^2 \right]}_{\text{(II.6)}}.
\end{aligned}$$

With (E.37) it is easy to see (together with the uniform bounded moments assumption) that (II.2)–(II.6) are of order $O_{\mathbb{P}}(n^{-1}\sum_i \pi_{ii}^2) = O_{\mathbb{P}}(k/n)$, hence asymptotically negligible. As for (II.1), note that

$$\begin{aligned}
(\text{II.1}) = &- \frac{1}{n} \sum_{\substack{i,j,j' \\ \text{distinct}}} \pi_{ij}^2 \pi_{ij'}^2 \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \varepsilon_j^2 \right]^\top \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \varepsilon_{j'}^2 \right] - \frac{2}{n} \sum_{\substack{i,i',j \\ \text{distinct}}} \pi_{ij}^2 \pi_{ii'}^2 \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \varepsilon_j^2 \right]^\top \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_{i'} \varepsilon_i^2 \right] \\
&- \frac{1}{n} \sum_{\substack{i,i',j \\ \text{distinct}}} \pi_{ij}^2 \pi_{ij'}^2 \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \varepsilon_j^2 \right]^\top \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_{i'} \varepsilon_j^2 \right] - \frac{1}{n} \sum_{i,j,i\neq j} \pi_{ij}^4 \left(\mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \varepsilon_j^2 \right] \right)^2 - \frac{1}{n} \sum_{i,j,i\neq j} \pi_{ij}^4 \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i \varepsilon_j^2 \right]^\top \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_j \varepsilon_i^2 \right].
\end{aligned}$$

Therefore we have (II.1) is of order $O_{\mathbb{P}}(n^{-1}\sum_i \pi_{ii}^2) = O_{\mathbb{P}}(k/n)$.

Variance, Term (III)

Next we consider (III), and still (E.38) implies

$$\mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\left(\frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \ddot{\mathbf{m}}_i \pi_{ij} \pi_{ii} \varepsilon_j \varepsilon_i \right)^2 \right] = \frac{4}{n} \sum_{i,j,\text{distinct}} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[|\ddot{\mathbf{m}}_i|^2 \pi_{ij}^2 \pi_{ii}^2 \varepsilon_i^2 \varepsilon_j^2 \right] + \frac{8}{n} \sum_{i,j,\text{distinct}} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_j \pi_{ij}^2 \pi_{ii} \pi_{jj} \varepsilon_i^2 \varepsilon_j^2 \right],$$

where the two terms are denoted by (III.1) and (III.2), respectively. For (III.1) it is bounded by

$$|(\text{III.1})| \asymp \frac{1}{n} \sum_{i} \pi_{ii}^2 \sum_{j\neq i} \pi_{ij}^2 = \frac{1}{n} \sum_{i} \pi_{ii}^3,$$

which is bounded by k/n due to (E.36). Similarly

$$|(\text{III.2})| \precsim \frac{1}{n}\sum_{i,j}\pi_{ii}\pi_{jj}\pi_{ij}^2 \leq \frac{1}{n}\sum_{i,j}\pi_{jj}\pi_{ij}^2 = O(k/n),$$

due to (E.36) and $\pi_{ii} \leq 1$.

Variance, Term (IV)

Finally we consider (IV), and the variance is

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2\right)^2\right] - \left(\frac{1}{\sqrt{n}}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2\right]\right)^2 \\
&= \frac{1}{n}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\mathsf{T}\ddot{\mathbf{m}}_j\pi_{ii}^2\pi_{jj}^2\varepsilon_i^2\varepsilon_j^2\right] - \left(\frac{1}{\sqrt{n}}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2\right]\right)^2 + \frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[|\ddot{\mathbf{m}}_i|^2\,\pi_{ii}^4\varepsilon_i^4\right].
\end{aligned}$$

And both terms are bounded by $O(k/n)$.

The last step is to show that one can essentially replace $\tilde{\mu}_i$ by μ_i in (E.10). This is trivial due to Assumption A.1(7), and the consistency assumption A.2(1). ■

SA-9.7 Theorem SA.5

By the condition $k = O(\sqrt{n})$, all terms of order $O_\mathbb{P}(\sqrt{k/n})$ can be ignored asymptotically. Also the bias term has order $\boldsymbol{\mathcal{B}} = O_\mathbb{P}(k/\sqrt{n}) = O_\mathbb{P}(1)$. In particular, both (E.9) and (E.10) are of order $O_\mathbb{P}(1)$. By Assumption A.2(1), the remainder term in the quadratic expansion (after (E.10)) has the order $o_\mathbb{P}(|(\text{E.10})|)$, which is negligible. ■

SA-9.8 Theorem SA.6

We first make the following decomposition:

$$\tilde{\boldsymbol{\Psi}}_1 = \mathbb{E}[\bar{\boldsymbol{\Psi}}_1|\mathbf{Z}], \qquad \tilde{\boldsymbol{\Psi}}_2 = \bar{\boldsymbol{\Psi}}_1 - \mathbb{E}[\bar{\boldsymbol{\Psi}}_1|\mathbf{Z}] + \bar{\boldsymbol{\Psi}}_2.$$

Then note that $\tilde{\boldsymbol{\Psi}}_1$ is mean zero, and $\tilde{\boldsymbol{\Psi}}_2$ is conditionally mean zero (on $\mathbf{Z}$). One special case is that $\tilde{\boldsymbol{\Psi}}_1 = 0$ almost surely, which will happen if the moment condition for the second step is actually a conditional moment restriction. In what follows, we assume $\tilde{\boldsymbol{\Psi}}_1$ is nondegenerate.

By the usual central limit theorem, one has

$$\left(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1]\right)^{-1/2}\tilde{\boldsymbol{\Psi}}_1 \rightsquigarrow \mathcal{N}(\mathbf{0},\mathbf{I}).$$

Next we consider the large sample distribution of $\tilde{\boldsymbol{\Psi}}_2$, which requires triangular array type argument. Let $\boldsymbol{\alpha}$ be a generic vector, and consider

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(a_i+b_i)^2\mathbb{1}\left[|a_i+b_i| > 2\varepsilon\sqrt{n}\right]\right],$$

where

$$a_i = \boldsymbol{\alpha}^\mathsf{T}\Big(\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) - \mathbb{E}[\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\mathbf{z}_i]\Big), \qquad b_i = \boldsymbol{\alpha}^\mathsf{T}\Big(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j,\mu_j,\boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij}\Big)\varepsilon_i.$$

Note that

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(a_i+b_i)^2\mathbb{1}\left[|a_i+b_i| > 2\varepsilon\sqrt{n}\right]\right] \precsim \frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(a_i^2+b_i^2\right)\left(\mathbb{1}\left[|a_i| > \varepsilon\sqrt{n}\right] + \mathbb{1}\left[|b_i| > \varepsilon\sqrt{n}\right]\right)\right],$$

which is a sum of four terms.

The first case is the easiest:

$$\mathbb{E}\left|\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right]\right| = \frac{1}{n}\sum_i \mathbb{E}\left[a_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right] \to 0,$$

where the first equality is true since the summands are nonnegative, and the last line comes from the i.i.d.ness of a_i. Therefore

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right] = o_{\mathbb{P}}(1).$$

For future reference, define $\tilde{b}_i = \boldsymbol{\alpha}^{\mathsf{T}}\left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j,\mu_j,\boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij}\right)$. Then the second case becomes (where we used the union bound)

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|b_i|>\varepsilon\sqrt{n}\right]\right] \le \frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\tilde{b}_i|>\varepsilon\sqrt{n}/\log(n)\right]\right] + \frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\varepsilon_i|>\log(n)\right]\right].$$

the last term in the above display is $o_{\mathbb{P}}(1)$ since it has expectation (note that it is nonnegative)

$$\begin{aligned}&\lim_n \mathbb{E}\left[\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\varepsilon_i|>\log(n)\right]\right]\right] = \lim_n \mathbb{E}\left[a_i^2\mathbb{1}\left[|\varepsilon_i|>\log(n)\right]\right]\\ &= \mathbb{E}\left[a_i^2\lim_n \mathbb{1}\left[|\varepsilon_i|>\log(n)\right]\right] = 0,\end{aligned}$$

and interchanging limit and expectation is justified by dominated convergence, and the fact that $\mathbb{E}[a_i^2]<\infty$. The other terms is handled by the following:

$$\begin{aligned}&\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\tilde{b}_i|>\varepsilon\sqrt{n}/\log(n)\right]\right] = \frac{1}{n}\sum_i \mathbb{1}\left[|\tilde{b}_i|>\varepsilon\sqrt{n}/\log(n)\right]\mathbb{E}_{[\cdot|\mathbf{Z}]}[a_i^2]\\ &\precsim \frac{1}{n}\sum_i \mathbb{1}\left[|\tilde{b}_i|>\varepsilon\sqrt{n}/\log(n)\right].\end{aligned}$$

The first line comes from the fact that $\tilde{b}_i$ is constant after conditioning on $\mathbf{Z}$, and the second line is true since $\mathbb{E}_{[\cdot|\mathbf{Z}]}[a_i^2]$ is bounded. We show it is $o_{\mathbb{P}}(1)$ again by taking expectation, and the fact that $\tilde{b}_i$ is the projection of random variable with finite expectation.

The next case is again very simple:

$$\begin{aligned}&\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[b_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right] \precsim \frac{1}{n}\sum_i \tilde{b}_i^2\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\varepsilon_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right]\\ &\le \left(\max_{1\le i\le n}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\varepsilon_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right]\right)\frac{1}{n}\sum_i \tilde{b}_i^2 \precsim \left(\max_{1\le i\le n}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\varepsilon_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right]\right)\to 0.\end{aligned}$$

The first inequality comes from the definition of $\tilde{b}_i$; the second is Hölder's inequality; the third inequality uses the fact $\sum_i \tilde{b}_i^2 = O(n)$; and the final inequality is true since we assumed bounded conditional moment.

Finally, the last case is

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[b_i^2\mathbb{1}\left[|b_i|>\varepsilon\sqrt{n}\right]\right] \precsim \frac{1}{n}\sum_i \tilde{b}_i^2\mathbb{1}\left[|\tilde{b}_i|>\varepsilon\sqrt{n}/\log(n)\right] + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1),$$

since $\tilde{b}_i$ comes from projecting a bounded sequence.

To summarize, we have the following two convergence results: (1) $\tilde{\boldsymbol{\Psi}}_1$ converges unconditionally to a multivariate normal distribution; and (2) conditional on $\mathbf{Z}$, $\tilde{\boldsymbol{\Psi}}_2$ converges to a multivariate normal distribution (more precisely, conditional on $\mathbf{Z}$ the distribution function of $\tilde{\boldsymbol{\Psi}}_2$ converges to that of a multivariate normal in probability). The following remark shows how joint convergence can be established (not that it is not true in general that one can conclude joint convergence from marginal convergence)

Remark (From marginal convergence to joint convergence)**.** Here we consider one special case where it is possible to

deduce joint convergence from marginal convergence. Assume $X_n \rightsquigarrow \mathcal{N}(0,1)$ and $Y_n|Z_n \rightsquigarrow_{\mathbb{P}} \mathcal{N}(0,1)$, and $X_n \in \sigma(Z_n)$, where $Y_n|Z_n \rightsquigarrow_{\mathbb{P}} \mathcal{N}(0,1)$. Then, $[X_n, Y_n]^{\mathsf{T}} \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$.

This follows because

$$\begin{aligned}\mathbb{P}\Big[X_n \le x, Y_n \le y\Big] &= \mathbb{E}\Big[\mathbb{1}[X_n \le x]\mathbb{P}[Y_n \le y|Z_n]\Big] \\ &= \mathbb{E}\Big[\mathbb{1}[X_n \le x]\Big(\mathbb{P}[Y_n \le y|Z_n] - \Phi(y)\Big)\Big] + \mathbb{P}\Big[X_n \le x\Big]\Phi(y) \\ &\to \Phi(x)\Phi(y),\end{aligned}$$

using the dominated convergence theorem and the assumption that $\mathbb{P}[Y_n \le y|Z_n] \to_{\mathbb{P}} \Phi(y)$. ⌟

Hence we are able to show

$$\begin{bmatrix} \Big(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1]\Big)^{-1/2}\tilde{\boldsymbol{\Psi}}_1 \\ \Big(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_2|\mathbf{Z}]\Big)^{-1/2}\tilde{\boldsymbol{\Psi}}_2 \end{bmatrix} \rightsquigarrow \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\right),$$

and the desired result follows by considering the linear combination

$$\Big(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1] + \mathbb{V}[\tilde{\boldsymbol{\Psi}}_2|\mathbf{Z}]\Big)^{-1/2}\left[\Big(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1]\Big)^{1/2}, \quad \Big(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_2|\mathbf{Z}]\Big)^{1/2}\right].$$

■

SA-9.9 Additional Details of Section SA-4.3

Given the sample estimating equation,

$$\mathbf{0} = \sum_i \Big(\sum_j q_{ij}\mathbf{f}(\mathbf{x}_j, \hat{\mu}_j, \hat{\boldsymbol{\theta}})\Big)\Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\Big),$$

Taylor expansion gives

$$\begin{aligned}\mathbf{0} = &\sum_i \Big(\sum_j q_{ij}\mathbf{f}(\mathbf{x}_j, \hat{\mu}_j, \boldsymbol{\theta}_0)\Big)\Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big) \\ &+ \left[\sum_i \left[\Big(\sum_j q_{ij}\mathbf{F}(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}})\Big)\Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}})\Big) - \Big(\sum_j q_{ij}\mathbf{f}(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}})\Big)\mathbf{f}(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}})^{\mathsf{T}}\right]\right]\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big),\end{aligned}$$

where $\check{\boldsymbol{\theta}}$ is some convex combination of $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$.

SA-9.9.1 Linearization

Consider the following:

$$\begin{aligned}&\left|\frac{1}{n}\sum_i \Big(\sum_j q_{ij}\mathbf{F}(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}})\Big)\Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}})\Big) - \frac{1}{n}\sum_i \Big(\sum_j q_{ij}\mathbf{F}_j\Big)\Big(y_i - f_i\Big)\right| \\ &\le \underbrace{\left|\frac{1}{n}\sum_i \Big(\sum_j q_{ij}(\mathbf{F}(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}}) - \mathbf{F}_j\Big)\Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}})\Big)\right|}_{\text{(I)}} + \underbrace{\left|\frac{1}{n}\sum_i \Big(\sum_j q_{ij}\mathbf{F}_j\Big)\Big(f(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}}) - f_i\Big)\right|}_{\text{(II)}}.\end{aligned}$$

Then

$$\begin{aligned}|\text{(II)}| &\le \left|\frac{1}{n}\sum_i \Big(\sum_j q_{ij}\mathbf{F}_j\Big)\mathcal{H}_i^{\alpha,\delta}(f)\Big(|\hat{\mu}_i - \mu_i| + |\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|\Big)^{\alpha}\right| \\ &\le \frac{o_{\mathbb{P}}(1)}{n}\sum_i \left|\Big(\sum_j q_{ij}\mathbf{F}_j\Big)\mathcal{H}_i^{\alpha,\delta}(f)\right| \le o_{\mathbb{P}}(1)\sqrt{\frac{1}{n}\sum_i |\mathbf{F}_i|^2}\sqrt{\frac{1}{n}\sum_i |\mathcal{H}_i^{\alpha,\delta}(f)|^2} = o_{\mathbb{P}}(1),\end{aligned}$$

where we used the assumption that $|\mathbf{F}_i|$ and $\mathcal{H}_i^{\alpha,\delta}(f)$ belong to BM_2. For (I),

$$
\begin{aligned}
|(\mathrm{I})| &= \left| \frac{1}{n} \sum_i \Big(\sum_j q_{ij}(y_j - f(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}})) \Big) \Big(\mathbf{F}(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}}) - \mathbf{F}_i \Big) \right| \\
&\leq \frac{op(1)}{n} \sum_i \left| \Big(\sum_j q_{ij}(y_j - f(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}})) \Big) \mathcal{H}_i^{\alpha,\delta}(\mathbf{F}) \right| \leq o_{\mathbb{P}}(1) \sqrt{\frac{1}{n} \sum_i |y_i - f(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}})|^2} \sqrt{\frac{1}{n} \sum_i |\mathcal{H}_i^{\alpha,\delta}(\mathbf{F})|^2} \\
&\leq o_{\mathbb{P}}(1) \sqrt{\frac{1}{n} \sum_i \Big(|y_i - f_i|^2 + 2|y_i - f_i| \mathcal{H}_i^{\alpha,\delta}(f) + |\mathcal{H}_i^{\alpha,\delta}(f)|^2 \Big)} \sqrt{\frac{1}{n} \sum_i |\mathcal{H}_i^{\alpha,\delta}(\mathbf{F})|^2} = o_{\mathbb{P}}(1),
\end{aligned}
$$

where we assumed that u_i, $\mathcal{H}_i^{\alpha,\delta}(f)$ and $\mathcal{H}_i^{\alpha,\delta}(\mathbf{F})$ belong to BM_2. Then (recall that $u_i = y_i - f_i$)

$$
\mathbb{E}\left[\frac{1}{n} \sum_i \Big(\sum_j q_{ij} \mathbf{F}_j \Big) u_i \,\middle|\, \mathbf{X}, \mathbf{Z} \right] = \mathbf{0},
$$

and the conditional variance is

$$
\left| \mathbb{V}\left[\frac{1}{n} \sum_i \Big(\sum_j q_{ij} \mathbf{F}_j \Big) u_i \,\middle|\, \mathbf{X}, \mathbf{Z} \right] \right| \precsim \frac{1}{n^2} \sum_i \Big| \sum_j q_{ij} \mathbf{F}_j \Big|^2 \leq \frac{1}{n^2} \sum_i |\mathbf{F}_i|^2 = o_{\mathbb{P}}(1),
$$

where we used the assumption that $\mathbb{V}[u_i|\mathbf{x}_i, \mathbf{z}_i]$ is uniformly bounded and that $|\mathbf{F}_i| \in \mathsf{BM}_2$. Therefore we have

$$
\frac{1}{n} \sum_i \Big(\sum_j q_{ij} \mathbf{F}(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}}) \Big) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}}) \Big) = o_{\mathbb{P}}(1).
$$

Using similar technique, we can show that, provided that $|_{\mathbf{i}}|$ and $\mathcal{H}_i^{\alpha,\delta}()$ are in BM_2, one has

$$
\begin{aligned}
&\frac{1}{n} \sum_i \Big(\sum_j q_{ij} \mathbf{f}(\mathbf{x}_j, \hat{\mu}_j, \check{\boldsymbol{\theta}}) \Big) \mathbf{f}(\mathbf{x}_i, \hat{\mu}_i, \check{\boldsymbol{\theta}})^{\mathsf{T}} = \frac{1}{n} \sum_i \Big(\sum_j q_{ij} \mathbf{f}_j \Big) \mathbf{f}_i^{\mathsf{T}} + o_{\mathbb{P}}(1) \\
&= \frac{1}{n} \sum_i \Big(\sum_j q_{ij} (\mathbf{f}_j - \mathbb{E}[\mathbf{f}_j|\mathbf{z}_j]) \Big) \mathbf{f}_i^{\mathsf{T}} + \frac{1}{n} \sum_i \Big(\sum_j q_{ij} \mathbb{E}[\mathbf{f}_j|\mathbf{z}_j] \Big) \mathbf{f}_i^{\mathsf{T}} + o_{\mathbb{P}}(1) \\
&= \frac{1}{n} \sum_i (\mathbf{f}_i - \mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]) \mathbf{f}_i^{\mathsf{T}} + o_{\mathbb{P}}(1) \to_{\mathbb{P}} \mathbb{E}\mathbb{V}[\mathbf{f}_i|\mathbf{z}_i],
\end{aligned}
$$

where for the last line, we used the assumption that the conditional expectation $\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]$ can be approximated by some linear span of $\mathbf{z}_i$, relying on the same argument used in Lemma SA.3. As a result, we established that

$$
\begin{aligned}
\sqrt{n}\Big(\mathbb{E}\mathbb{V}[\mathbf{f}_i|\mathbf{z}_i] \Big) \Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \Big) &= \left[\frac{1}{\sqrt{n}} \sum_i \Big(\sum_j q_{ij} \mathbf{f}(\mathbf{x}_j, \hat{\mu}_j, \boldsymbol{\theta}_0) \Big) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) \right] + o_{\mathbb{P}}(1) \\
&= \left[\frac{1}{\sqrt{n}} \sum_i \mathbf{f}(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) \right] \qquad \text{(III)} \\
&\quad - \left[\frac{1}{\sqrt{n}} \sum_i \Big(\sum_j \pi_{ij} \mathbf{f}(\mathbf{x}_j, \hat{\mu}_j, \boldsymbol{\theta}_0) \Big) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) \right] \qquad \text{(IV)} \\
&\quad + o_{\mathbb{P}}(1).
\end{aligned}
$$

SA-9.9.2 Term (III)

Note that (III) can be handled by employing the same techniques and regularity conditions employed in Theorem SA.5 (and the corresponding lemmas), by matching the notation: $\mathbf{m}_i = \mathbf{f}_i(y_i - f_i)$, hence we do not repeat the derivation. The following holds:

$$
\begin{aligned}
\frac{1}{\sqrt{n}} \sum_i \mathbf{f}(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) &= \frac{1}{\sqrt{n}} \sum_i \Big(\mathbf{f}_i u_i - \mathbb{E}[\mathbf{f}_i \dot{f}_i|\mathbf{z}_i] \varepsilon_i \Big) + \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\dot{\mathbf{f}}_i u_i \varepsilon_i - \mathbf{f}_i \dot{f}_i \varepsilon_i |\mathbf{z}_i] \pi_{ii} \\
&\quad + \frac{1}{\sqrt{n}} \sum_i \Big(-\frac{1}{2} \mathbb{E}[\mathbf{f}_i \ddot{f}_i|\mathbf{z}_i] - \mathbb{E}[\dot{\mathbf{f}}_i \dot{f}_i|\mathbf{z}_i] \Big) \mathbb{E}[\varepsilon_j^2|\mathbf{z}_j] \pi_{ij}^2 + o_{\mathbb{P}}(1).
\end{aligned}
$$

SA-9.9.3 Term (IV)

Next we expand (IV) with respect to estimated $\hat{\mu}$.

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_i \Big(\sum_j \pi_{ij}\mathbf{f}(\mathbf{x}_j,\hat{\mu}_j,\boldsymbol{\theta}_0)\Big)\Big(y_i - f(\mathbf{x}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\Big)\\
&= \frac{1}{\sqrt{n}}\sum_i \Big(\sum_j \pi_{ij}\mathbf{f}_j\Big)u_i && \text{(i)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \left[\dot{\mathbf{f}}_i\Big(\sum_j \pi_{ij}u_j\Big) - \Big(\sum_j \pi_{ij}\mathbf{f}_j\Big)\dot{f}_i\right]\Big(\hat{\mu}_i - \mu_i\Big) && \text{(ii)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \frac{1}{2}\left[\ddot{\mathbf{f}}_i\Big(\sum_j \pi_{ij}u_j\Big) - \Big(\sum_j \pi_{ij}\mathbf{f}_j\Big)\ddot{f}_i\right]\Big(\hat{\mu}_i - \mu_i\Big)^2 && \text{(iii)}\\
&+ \frac{1}{\sqrt{n}}\sum_{i,j} \dot{\mathbf{f}}_j\dot{f}_i\Big(\hat{\mu}_i - \mu_i\Big)\Big(\hat{\mu}_j - \mu_j\Big)\pi_{ij} && \text{(iv)}\\
&+ o_{\mathbb{P}}(1),
\end{aligned}$$

where again the remainder term (higher order expansion) is negligible since it involves cubics of $\hat{\mu}_i - \mu_i$.

SA-9.9.4 Term (i)

Term (i) has the following further expansion:

$$\text{(i)} = \frac{1}{\sqrt{n}}\sum_{i,j}\mathbb{E}[\mathbf{f}_j|\mathbf{z}_j]u_i\pi_{ij} + \frac{1}{\sqrt{n}}\sum_{i,j}(\mathbf{f}_j - \mathbb{E}[\mathbf{f}_j|\mathbf{z}_j])u_i\pi_{ij} = \frac{1}{\sqrt{n}}\sum_i \mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]u_i + o_{\mathbb{P}}(1),$$

with the same argument used in Lemma SA.3.

SA-9.9.5 Term (ii)

For (ii), we consider the following:

$$\text{(ii)} = \underbrace{\frac{1}{\sqrt{n}}\sum_{i,j}\dot{\mathbf{f}}_i u_j\pi_{ij}\Big(\hat{\mu}_i - \mu_i\Big)}_{\text{(ii.1)}} \underbrace{- \frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{f}_j\dot{f}_i\pi_{ij}\Big(\hat{\mu}_i - \mu_i\Big)}_{\text{(ii.2)}}.$$

We note that (ii.1) has essentially a quadratic form, hence the same technique of Lemma SA.4 can be applied with the obvious notation matching $\ddot{\mathbf{m}}_i = \dot{\mathbf{f}}_i$, which implies the following:

$$\begin{aligned}
\text{(ii.1)} &= \frac{1}{\sqrt{n}}\sum_{i,j,k}\dot{\mathbf{f}}_i u_j\varepsilon_k\pi_{ij}\pi_{ik} - \frac{1}{\sqrt{n}}\sum_{i,j,k}\dot{\mathbf{f}}_i u_j\eta_k\pi_{ij}q_{ik}\\
&= \frac{1}{\sqrt{n}}\sum_{i,j,k}\dot{\mathbf{f}}_i u_j\varepsilon_k\pi_{ij}\pi_{ik} + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}}\sum_{i,j}\mathbb{E}[\dot{\mathbf{f}}_i u_j\varepsilon_j|\mathbf{z}_i,\mathbf{z}_j]\pi_{ij}^2 + o_{\mathbb{P}}(1),
\end{aligned}$$

so that this term has bias contribution. Note that the term involving the approximation error η_i is negligible, since its conditional mean has order:

$$\left|\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j,k}\dot{\mathbf{f}}_i u_j\eta_k\pi_{ij}q_{ik}\right]\right| \precsim \frac{1}{\sqrt{n}}\sum_i \pi_{ii}\sum_j q_{ij}\eta_j,$$

which is negligible provided that $\sqrt{n}\mathbb{E}[\eta_i^2] = o(1)$ and $|\dot{\mathbf{f}}_i u_i| \in \mathsf{BCM}_2$. The conditional variance has order

$$\left|\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j,k}\dot{\mathbf{f}}_i u_j\eta_k\pi_{ij}q_{ik}\right]\right| \precsim \frac{1}{n}\sum_i \check{\eta}_i^2\mathbb{E}\left[|\dot{\mathbf{f}}_i|^2\Big(\sum_j \pi_{ij}u_j\Big)^2\middle|\mathbf{z}_i\right] \qquad (\text{where } \check{\eta}_i = \sum_j q_{ij}\eta_j)$$

$$\asymp \frac{1}{n}\sum_{i,j}\tilde{\eta}_i^2\pi_{ij}^2 = o_{\mathbb{P}}(1).$$

Next

$$\text{(ii.2)} = \underbrace{-\frac{1}{\sqrt{n}}\sum_{i,j}\mathbb{E}[\mathbf{f}_j|\mathbf{z}_j]\dot{f}_i\pi_{ij}(\hat{\mu}_i-\mu_i)}_{\text{(ii.2.1)}}\underbrace{-\frac{1}{\sqrt{n}}\sum_{i,j}(\mathbf{f}_j-\mathbb{E}[\mathbf{f}_j|\mathbf{z}_j])\dot{f}_i\pi_{ij}(\hat{\mu}_i-\mu_i)}_{\text{(ii.2.2)}}.$$

Note that (ii.2.2) can be handled in the same way as (II.1). That is, by employing Lemma SA.4 with $\ddot{\mathbf{m}}_i = \dot{f}_i$:

$$\text{(ii.2.2)} = -\frac{1}{\sqrt{n}}\sum_{i,j,k}(\mathbf{f}_j-\mathbb{E}[\mathbf{f}_j|\mathbf{z}_j])\dot{f}_i\varepsilon_k\pi_{ij}\pi_{ik}+o_{\mathbb{P}}(1) = -\frac{1}{\sqrt{n}}\sum_{i,j}\mathbb{E}[\dot{f}_i\mathbf{f}_j\varepsilon_j|\mathbf{z}_i,\mathbf{z}_j]\pi_{ij}^2+o_{\mathbb{P}}(1),$$

which has bias contribution. For (ii.2.1), we assume the approximation error satisfies: $\sqrt{n}\mathbb{E}[|\boldsymbol{\zeta}_{\mathbf{f},i}|^2] = o(1)$, where $\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i] = \mathbf{A}\mathbf{z}_i + \boldsymbol{\zeta}_{\mathbf{f},i}$ and $\mathbb{E}[\boldsymbol{\zeta}_{\mathbf{f},i}\mathbf{z}_i^{\mathsf{T}}] = \mathbf{0}$. Then Cauchy-Schwarz implies

$$\text{(ii.2.1)} = -\frac{1}{\sqrt{n}}\sum_{i}\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]\dot{f}_i(\hat{\mu}_i-\mu_i)+o_{\mathbb{P}}(1).$$

We emphasize that the above is not necessary, but greatly simplifies the final formula. Now we can employ the same technique of Lemma SA.3, which shows that (ii.2.1) has both variance and bias contribution:

$$\text{(ii.2.1)} = -\frac{1}{\sqrt{n}}\sum_{i,j}\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]\mathbb{E}[\dot{f}_i|\mathbf{z}_i]\varepsilon_i - \frac{1}{\sqrt{n}}\sum_{i}\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]\mathbb{E}[\dot{f}_i\varepsilon_i|\mathbf{z}_i]\pi_{ii}+o_{\mathbb{P}}(1),$$

where again we assume $\mathbb{E}[\dot{f}_i|\mathbf{z}_i]$ can be approximated by linear span of $\mathbf{z}_i$.

SA-9.9.6 Term (iii) and (iv)

The same line of reasoning can be employed to show that both (iii) and (iv) contributes to the asymptotic bias, by conditional expectation and variance calculation. To avoid the tedious and lengthy derivations, we only calculate the conditional expectation.

Term (iii) is again split as

$$\text{(iii)} = \underbrace{-\frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\ddot{\mathbf{f}}_i u_j\pi_{ij}\left(\hat{\mu}_i-\mu_i\right)^2}_{\text{(iii.1)}} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\mathbf{f}_j\ddot{f}_i\pi_{ij}\left(\hat{\mu}_i-\mu_i\right)^2}_{\text{(iii.2)}},$$

First,

$$\text{(iii.1)} = -\frac{1}{\sqrt{n}}\sum_{i,j,k,\ell}\frac{1}{2}\ddot{\mathbf{f}}_i u_j\varepsilon_k\varepsilon_\ell\pi_{ij}\pi_{ik}\pi_{i\ell} = -\frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\mathbb{E}[\ddot{\mathbf{f}}_i u_j\varepsilon_j^2|\mathbf{z}_j]\pi_{ij}^3+o_{\mathbb{P}}(1),$$

which is a bias contribution. Next,

$$\begin{aligned}
\text{(iii.2)} &= \frac{1}{\sqrt{n}}\sum_{i,j,k,\ell}\frac{1}{2}\mathbf{f}_j\ddot{f}_i\varepsilon_k\varepsilon_\ell\pi_{ij}\pi_{ik}\pi_{i\ell}\\
&= \frac{1}{\sqrt{n}}\sum_{i,k,\ell}\frac{1}{2}\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]\ddot{f}_i\varepsilon_k\varepsilon_\ell\pi_{ik}\pi_{i\ell} + \frac{1}{\sqrt{n}}\sum_{i,j,k,\ell}\frac{1}{2}(\mathbf{f}_j-\mathbb{E}[\mathbf{f}_j|\mathbf{z}_j])\ddot{f}_i\varepsilon_k\varepsilon_\ell\pi_{ij}\pi_{ik}\pi_{i\ell}\\
&= \frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]\mathbb{E}[\ddot{f}_i\varepsilon_j^2|\mathbf{z}_i,\mathbf{z}_j]\pi_{ij}^2 + \frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\mathbb{E}[\ddot{f}_i\mathbf{f}_j\varepsilon_j^2|\mathbf{z}_i,\mathbf{z}_j]\pi_{ij}^3+o_{\mathbb{P}}(1),
\end{aligned}$$

which is again a bias contribution.

Finally for (iv),

$$\text{(iv)} = \frac{1}{\sqrt{n}}\sum_{i,j,k,\ell}\dot{\mathbf{f}}_j\dot{f}_i\varepsilon_k\varepsilon_\ell\pi_{ij}\pi_{ik}\pi_{j\ell}$$

$$= \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}[\dot{f}_i \dot{\mathbf{f}}_j \varepsilon_i \varepsilon_j | \mathbf{z}_i, \mathbf{z}_j] \pi_{ij} \pi_{ii} \pi_{jj} + \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}[\dot{f}_i \dot{\mathbf{f}}_j \varepsilon_i \varepsilon_j | \mathbf{z}_i, \mathbf{z}_j] \pi_{ij}^3 + \frac{1}{\sqrt{n}} \sum_{i,j,k} [\dot{f}_i \dot{\mathbf{f}}_j \varepsilon_k^2 | \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k] \pi_{ij} \pi_{ik} \pi_{jk} + o_{\mathbb{P}}(1)$$
$$= \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}[\dot{f}_i \dot{\mathbf{f}}_j \varepsilon_i \varepsilon_j | \mathbf{z}_i, \mathbf{z}_j] \pi_{ij}^3 + \frac{1}{\sqrt{n}} \sum_{i,j,k} \mathbb{E}[\dot{f}_i \dot{\mathbf{f}}_j \varepsilon_k^2 | \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k] \pi_{ij} \pi_{ik} \pi_{jk} + o_{\mathbb{P}}(1).$$

Here we make a complementary calculation:

$$\sum_{i,j} |\pi_{ij} \pi_{ii} \pi_{jj}| \leq \left(\sum_{i,j} \pi_{ii} \right)^{1/2} \left(\sum_i \left(\sum_j \pi_{ij} \pi_{jj} \right)^2 \right)^{1/2} \qquad \text{(Cauchy-Schwarz)}$$
$$= \left(\sum_i \pi_{ii} \right)^{1/2} \left(\sum_{i,j,k} \pi_{ij} \pi_{jj} \pi_{ik} \pi_{kk} \right)^{1/2} = \left(\sum_i \pi_{ii} \right)^{1/2} \left(\sum_{i,j} \pi_{ij} \pi_{ii} \pi_{jj} \right)^{1/2},$$

which shows that the bias contribution from (iv) is indeed of order $k/\sqrt{n}$. ■

SA-9.10 Proposition SA.7

Here we do some calculations. Note that in this example, $\mathbf{w}_i = [\mathbf{Y}_i^\intercal, T_i, \mathbf{X}_i^\intercal]^\intercal$, $\mu_i = P_i$ and $\mathbf{z}_i = \mathbf{Z}_i$, and

$$\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}) = \frac{T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta})}{P_i},$$

hence the two derivatives (with respect to $\mu_i = P_i$) are

$$\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = -\frac{T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i^2}, \qquad \frac{1}{2} \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = \frac{T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i^3}.$$

To compute the bias term, we need the following:

$$\mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \varepsilon_i | \mathbf{z}_i\right] = -\mathbb{E}\left[\frac{T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i^2} \varepsilon_i \middle| \mathbf{Z}_i \right] = -\mathbb{E}\left[\frac{T_i \mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i^2} \varepsilon_i \middle| \mathbf{Z}_i \right] = -\mathbf{g}_i \mathbb{E}\left[\frac{T_i}{P_i^2} \varepsilon_i \middle| \mathbf{Z}_i \right]$$
$$= -\mathbf{g}_i \frac{1 - P_i}{P_i},$$

which is $\mathbf{b}_{1,i}$. Similarly, one can show that

$$\mathbf{b}_{2,ij} = \mathbb{E}\left[\frac{1}{2} \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \varepsilon_j^2 \middle| \mathbf{z}_i, \mathbf{z}_j \right] = \mathbb{E}\left[\frac{T_i \mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i^3} \varepsilon_j^2 \middle| \mathbf{Z}_i, \mathbf{Z}_j \right]$$
$$= \mathbb{E}\left[\frac{T_i \mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0)}{P_i^3} \varepsilon_j^2 \middle| \mathbf{Z}_i, \mathbf{Z}_j \right] = \mathbf{g}_i \frac{\mathbb{E}[T_i \varepsilon_j^2 | \mathbf{Z}_i, \mathbf{Z}_j]}{P_i^3}.$$

Finally we consider the variance contribution, which utilizes

$$\mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_i, \boldsymbol{\theta}_0) | \mathbf{z}_j\right] = -\mathbb{E}\left[\frac{T_j \mathbf{h}(\mathbf{Y}_j, \mathbf{X}_j, \boldsymbol{\theta}_0)}{P_j^2} \middle| \mathbf{Z}_j \right] = -\mathbf{g}_j \frac{1}{P_j}.$$

■

SA-9.11 Proposition SA.9

$\hat{\theta}$ has the expansion

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \frac{1}{\sqrt{n}\mathbb{P}[T_i = 1]} \sum_i \left[\frac{T_i - \hat{P}_i}{1 - \hat{P}_i} \left(Y_i(t_2) - Y_i(t_1) \right) - T_i \theta_0 \right] + o_{\mathbb{P}}(1).$$

To calculate the bias and variance, note that the estimating equation depends on the first step through $(T_i - P_i)/(1 - P_i)$, hence it suffices to consider its derivatives with respect to P_i:

$$\frac{\partial}{\partial P_i}\left\{\frac{T_i - P_i}{1 - P_i}\right\} = \frac{T_i - 1}{(1 - P_i)^2}, \qquad \frac{1}{2}\frac{\partial^2}{\partial P_i^2}\left\{\frac{T_i - \hat{P}_i}{1 - \hat{P}_i}\right\} = \frac{T_i - 1}{(1 - P_i)^3}.$$

Therefore the first bias term is

$$\begin{aligned} b_{1,i} &= \mathbb{E}\left[\frac{T_i - 1}{(1 - P_i)^2}\Big(Y_i(t_2) - Y_i(t_1)\Big)(T_i - P_i)\,\Big|\,\mathbf{X}_i\right] = \mathbb{E}\left[\frac{T_i - 1}{1 - P_i}\Big(Y_i(t_2) - Y_i(t_1)\Big)(T_i - P_i)\,\Big|\,T_i = 0, \mathbf{X}_i\right] \\ &= \mathbb{E}\Big[\frac{P_i}{1 - P_i}\Big(Y_i(0, t_2) - Y_i(t_1)\Big)\Big|T_i = 0, \mathbf{X}_i\Big] = \frac{P_i}{1 - P_i}\mathbb{E}\Big[Y_i(0, t_2) - Y_i(t_1)\Big|T_i = 1, \mathbf{X}_i\Big], \end{aligned}$$

where the last line uses Assumption A.DiD(1). Note that the first bias term essentially reflects the trend component. The second bias term is

$$b_{2,ij} = \mathbb{E}\left[\frac{T_i - 1}{(1 - P_i)^3}\Big(Y_i(t_2) - Y_i(t_1)\Big)(T_j - P_j)^2\,\Big|\,\mathbf{X}_i, \mathbf{X}_j\right] = \mathbb{E}\left[-\frac{1}{(1 - P_i)^2}\Big(Y_i(t_2) - Y_i(t_1)\Big)(T_j - P_j)^2\,\Big|\,T_i = 0, \mathbf{X}_i, \mathbf{X}_j\right],$$

which gives

$$b_{2,ii} = -\frac{P_i^2}{(1 - P_i)^2}\mathbb{E}\left[Y_i(0, t_2) - Y_i(t_1)\middle|\, T_i = 1, \mathbf{X}_i\right],$$

or when $i \neq j$,

$$b_{2,ij} = -\frac{P_j(1 - P_j)}{(1 - P_i)^2}\mathbb{E}\left[Y_i(0, t_2) - Y_i(t_1)\middle|\, T_i = 1, \mathbf{X}_i\right]. \qquad (i \neq j)$$

And again, this depends on the trend component. To simplify, note that it is

$$b_{2,ij} = -\frac{\mathbb{E}[(T_j - P_j)^2 | T_i = 0, \mathbf{X}_i, \mathbf{X}_j]}{(1 - P_i)^2}\mathbb{E}\left[Y_i(0, t_2) - Y_i(t_1)\middle|\, T_i = 1, \mathbf{X}_i\right]$$

Finally, the variance contribution of the first step can be computed with the following:

$$\mathbb{E}\left[\frac{T_j - 1}{(1 - P_j)^2}\Big(Y_j(t_2) - Y_j(t_1)\Big)\,\Big|\,\mathbf{X}_j\right] = -\frac{1}{1 - P_j}\mathbb{E}\left[Y_j(0, t_2) - Y_j(t_1)\middle|\, T_j = 1, \mathbf{X}_j\right],$$

which gives

$$\bar{\Psi}_2 = -\frac{1}{\sqrt{n}\mathbb{P}[T_i = 1]}\sum_i\left[\sum_j \frac{1}{1 - P_j}\mathbb{E}\left[Y_j(0, t_2) - Y_j(t_1)\middle|\, T_j = 1, \mathbf{X}_j\right]\pi_{ij}\right]\varepsilon_i.$$

■

SA-9.12 Proposition SA.10

Since the estimating equation depends on the unobserved probability $\mu_i = P_i$ only through κ_i, we have the partial derivatives (with respect to $\mu_i = P_i$)

$$\dot{\kappa}_i = \frac{\partial}{\partial P_i}\kappa_i = -\frac{Y_i(1 - D_i)}{(1 - P_i)^2} + \frac{(1 - T_i)D_i}{P_i^2}, \quad \ddot{\kappa}_i = \frac{\partial^2}{\partial P_i^2}\kappa_i = -\frac{2T_i(1 - D_i)}{(1 - P_i)^3} - \frac{2(1 - T_i)D_i}{P_i^3},$$

hence

$$\begin{aligned} \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) &= \frac{\partial}{\partial \boldsymbol{\theta}}e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big)\left(-\frac{T_i(1 - D_i)}{(1 - P_i)^2} + \frac{(1 - T_i)D_i}{P_i^2}\right) \\ \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) &= \frac{\partial}{\partial \boldsymbol{\theta}}e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big)\left(-\frac{T_i(1 - D_i)}{(1 - P_i)^3} - \frac{(1 - T_i)D_i}{P_i^3}\right). \end{aligned}$$

To characterize the bias, one has to use more delicate arguments, and we do this for each term separately. Recall that $\varepsilon_i = D_i - P_i$, and by Assumption A.2(2), conditioning on $\mathbf{Z}_i$ alone will be asymptotically equivalent to conditioning

on both $\mathbf{Z}_i$ and P_i. For notational simplicity, define

$$e_{i,(\bullet)}(\boldsymbol{\theta}) = e(\mathbf{x}_i, T_i(\bullet), \boldsymbol{\theta}), \qquad \bullet = 0, 1,$$

for the two potential treatment status. Then observe that

$$\begin{aligned}
\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = &- \frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big)\frac{T_i(0)(1-D_i)}{(1-P_i)^2} \\
&+ \frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big)\frac{(1-T_i(1))D_i}{P_i^2}, \\
\frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = &- \frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big)\frac{T_i(0)(1-D_i)}{(1-P_i)^3} \\
&- \frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big)\frac{(1-T_i(1))D_i}{P_i^3}.
\end{aligned}$$

To understand the source of the bias, we first consider one piece:

$$\begin{aligned}
&\mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big)\frac{T_i(0)(1-D_i)}{(1-P_i)^2}\varepsilon_i \,\middle|\, \mathbf{Z}_i\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big)T_i(0) \,\middle|\, \mathbf{Z}_i\right]\frac{P_i}{1-P_i} \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big)T_i(0) \,\middle|\, \mathbf{Z}_i, T_i(0) = T_i(1)\right]\frac{P_i}{1-P_i} \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i] \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big)\frac{T_i(0)P_i}{1-P_i} \,\middle|\, \mathbf{Z}_i, T_i(0) = T_i(1)\right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i] \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big)\frac{T_i P_i}{1-P_i} \,\middle|\, \mathbf{Z}_i, T_i(0) = T_i(1)\right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i],
\end{aligned}$$

where the second and the fourth line follow from Assumption A.LARF(2), and the third lines uses the fact that there are no defiers, and for compliers, the conditional expectation is zero. Similarly, we can establish the following:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big)\frac{(1-T_i(1))D_i}{P_i^2}\varepsilon_i \,\middle|\, \mathbf{Z}_i\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big)(1-T_i(1)) \,\middle|\, \mathbf{Z}_i\right]\frac{1-P_i}{P_i} \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big)(1-T_i(1)) \,\middle|\, \mathbf{Z}_i, T_i(0) = T_i(1)\right]\frac{1-P_i}{P_i} \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i] \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big)\frac{(1-T_i(1))(1-P_i)}{P_i} \,\middle|\, \mathbf{Z}_i, T_i(0) = T_i(1)\right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i] \\
&= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big)\frac{(1-T_i)(1-P_i)}{P_i} \,\middle|\, \mathbf{Z}_i, T_i(0) = T_i(1)\right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i],
\end{aligned}$$

hence the first bias term takes the form:

$$\mathbf{b}_{1,i} = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big)\left(\frac{T_i P_i}{1-P_i} + \frac{(1-T_i)(1-P_i)}{P_i}\right) \,\middle|\, \mathbf{Z}_i, T_i(0) = T_i(1)\right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i].$$

For the other two cases, we use essentially the same technique:

$$\begin{aligned}
&\mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big)\frac{T_i(0)(1-D_i)}{(1-P_i)^3}\varepsilon_j^2 \,\middle|\, \mathbf{Z}_i, \mathbf{Z}_j\right] \\
&= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big)\frac{T_i(1-D_i)\varepsilon_j^2}{(1-P_i)^3} \,\middle|\, \mathbf{Z}_i, \mathbf{Z}_j, T_i(0) = T_i(1)\right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i],
\end{aligned}$$

and

$$\mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big)\frac{(1-T_i(1))D_i}{P_i^3}\varepsilon_j^2 \,\middle|\, \mathbf{Z}_i, \mathbf{Z}_j\right]$$

$$= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \frac{(1-T_i)D_i\varepsilon_j^2}{P_i^3} \,\Big|\, \mathbf{Z}_i, \mathbf{Z}_j, T_i(0) = T_i(1) \right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i],$$

which gives

$$\mathbf{b}_{2,ij} = \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \left(\frac{(1-T_i)D_i}{P_i^3} + \frac{T_i(1-D_i)}{(1-P_i)^3} \right) \varepsilon_j^2 \,\Big|\, \mathbf{Z}_i, \mathbf{Z}_j, T_i(0) = T_i(1) \right] \cdot \mathbb{P}[T_i(0) = T_i(1)|\mathbf{Z}_i].$$

For the variance, we need the following:

$$\mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big) \frac{T_i(0)(1-D_i)}{(1-P_i)^2} \,\Big|\, \mathbf{Z}_i \right] = \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(0)}(\boldsymbol{\theta}_0)\Big(Y_i(1) - e_{i,(0)}(\boldsymbol{\theta}_0)\Big) \frac{T_i(0)}{1-P_i} \,\Big|\, \mathbf{Z}_i \right]$$
$$= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \frac{T_i}{1-P_i} \,\Big|\, \mathbf{Z}_i, T_i(0) = T_i(1) \right] \cdot \mathbb{P}\left[T_i(0) = T_i(1)|\mathbf{Z}_i\right],$$

and

$$\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big) \frac{(1-T_i(1))D_i}{P_i^2} \,\Big|\, \mathbf{Z}_i \right] = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_{i,(1)}(\boldsymbol{\theta}_0)\Big(Y_i(0) - e_{i,(1)}(\boldsymbol{\theta}_0)\Big) \frac{1-T_i(1)}{P_i} \,\Big|\, \mathbf{Z}_i \right]$$
$$= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \frac{1-T_i}{P_i} \,\Big|\, \mathbf{Z}_i, T_i(0) = T_i(1) \right] \cdot \mathbb{P}\left[T_i(0) = T_i(1)|\mathbf{Z}_i\right],$$

hence

$$\mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{Z}_i\right] = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \left(\frac{1-T_i}{P_i} - \frac{T_i}{1-P_i} \right) \,\Big|\, \mathbf{Z}_i, T_i(0) = T_i(1) \right] \cdot \mathbb{P}\left[T_i(0) = T_i(1)|\mathbf{Z}_i\right],$$

which will be part of the asymptotic representation. ■

SA-9.13 Proposition SA.11

To match the notation used in the general result, note that $r_i = T_i$ and $\mu_i = P_i$, which we will use in the following calculations.

The derivation of the bias and variance is pretty straightforward. Note that

$$\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)\Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)\Big) = \frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big)$$

hence

$$\begin{aligned}
\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) &= \frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)\Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)\Big) - \frac{\partial}{\partial \boldsymbol{\theta}} e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \cdot \dot{e}(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) - \frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0) \cdot \dot{e}_i(\boldsymbol{\theta}_0) \\
\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) &= \frac{\partial}{\partial \boldsymbol{\theta}} \ddot{e}(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)\Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)\Big) - 2\frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \cdot \dot{e}(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \\
&\qquad - \frac{\partial}{\partial \boldsymbol{\theta}} e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \cdot \ddot{e}(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \ddot{e}_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) - 2\frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}_i(\boldsymbol{\theta}_0) \cdot \dot{e}_i(\boldsymbol{\theta}_0) - \frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0) \cdot \ddot{e}_i(\boldsymbol{\theta}_0).
\end{aligned}$$

Then we can compute the bias terms,

$$\begin{aligned}
\mathbf{b}_{1,i} &= \mathbb{E}\Big[\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) - \frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0) \cdot \dot{e}_i(\boldsymbol{\theta}_0)\Big]\varepsilon_i \Big| \mathbf{Z}_i\Big] = \mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}_i(\boldsymbol{\theta}_0) Y_i \varepsilon_i \Big| \mathbf{Z}_i\Big] \\
&= \mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}_i(\boldsymbol{\theta}_0) \left[T_i Y_i(1) + (1-T_i) Y_i(0)\right] \varepsilon_i \Big| \mathbf{Z}_i\Big],
\end{aligned}$$

from which one can get the desired result. Similarly we can derive the formula for $\mathbf{b}_{2,ij}$. Note that it suffices to consider the case $i \neq j$, as the terms involving π_{ii}^2 is asymptotically negligible:

$$\mathbf{b}_{2,ij} = \frac{1}{2}\mathbb{E}\Big[\Big[-2\frac{\partial}{\partial \boldsymbol{\theta}} \dot{e}_i(\boldsymbol{\theta}_0) \cdot \dot{e}_i(\boldsymbol{\theta}_0) - \frac{\partial}{\partial \boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0) \cdot \ddot{e}_i(\boldsymbol{\theta}_0)\Big]\varepsilon_j^2 \Big| \mathbf{Z}_i, \mathbf{Z}_j\Big]$$

$$= -\frac{1}{2}\left(2\frac{\partial}{\partial\boldsymbol{\theta}}\dot{e}_i(\boldsymbol{\theta}_0)\cdot\dot{e}_i(\boldsymbol{\theta}_0) + \frac{\partial}{\partial\boldsymbol{\theta}}e_i(\boldsymbol{\theta}_0)\cdot\ddot{e}_i(\boldsymbol{\theta}_0)\right)P_i(1-P_i).$$

Finally, one can also recover the variance term in a similar way. ■

SA-9.14 Proposition SA.12

To derive the bias and variance, we maintain the "dot" notation to denote the partial derivative with respect to μ_i, which has to be estimated in the first step. Then $\dot{\mathbf{X}}_i = -\mathbf{e}_2 = [0,-1]^\mathsf{T}$. Hence

$$\begin{aligned}\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) &= \frac{\partial}{\partial\mu_i}\left\{\mathbf{X}_iL'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big)\right\} = \dot{\mathbf{X}}_iL'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big) + \mathbf{X}_i\dot{\mathbf{X}}_i^\mathsf{T}\boldsymbol{\theta}_0L''(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big) - \mathbf{X}_i\dot{\mathbf{X}}_i^\mathsf{T}\boldsymbol{\theta}_0\\ &= -\mathbf{e}_2L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big) + \gamma_0\mathbf{X}_iL''(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big) - \gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2.\end{aligned}$$

Thanks to Assumption A.CF(1), the first bias component is

$$\mathbf{b}_{1,i} = \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_i|\mathbf{Z}_i\right] = -\mathbb{E}[\gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2\varepsilon_i|\mathbf{Z}_i],$$

and for the variance component $\bar{\boldsymbol{\Psi}}_2$,

$$\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\mathbf{Z}_i] = -\mathbb{E}[\gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2|\mathbf{Z}_i].$$

Using similar logic, it is not hard to show that the second bias term takes the form:

$$\begin{aligned}\mathbf{b}_{2,ij} &= \mathbb{E}\left[\frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right] = \mathbb{E}\left[\frac{1}{2}\frac{\partial}{\partial\mu_i}\left[-\mathbf{e}_2L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big)\right]\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right]\\ &\quad + \mathbb{E}\left[\frac{1}{2}\frac{\partial}{\partial\mu_i}\left[\gamma_0\mathbf{X}_iL''(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\Big)\right]\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right] + \mathbb{E}\left[\frac{1}{2}\frac{\partial}{\partial\mu_i}\left[-\gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2\right]\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right]\\ &= \mathbb{E}\left[\frac{\gamma_0}{2}\mathbf{e}_2L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)^2\varepsilon_j^2 - \frac{\gamma_0^2}{2}\mathbf{X}_iL''(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)L'(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\varepsilon_j^2 - \gamma_0^2\mathbf{X}_iL''(\mathbf{X}_i^\mathsf{T}\boldsymbol{\theta}_0)\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right].\end{aligned}$$

■

SA-9.15 Proposition SA.13

To simplify notation, we let $\mathbb{E}_{t_1}[\cdot] = \mathbb{E}[\cdot|I_{i,t_1},K_{i,t_1},A_{i,t_1}]$. Note that it is *not* the expectation conditional on the full time-t_1 information.

The first derivatives of $\mathbf{m}$ with respect to ν_{1i} and μ_{2i} are

$$\dot{\mathbf{m}}_1(\mathbf{w}_i,\nu_{1i}-z_{11i}\gamma,\mu_{2i},\gamma,\boldsymbol{\theta}) = \begin{bmatrix}K_{i,t_1}g_{22,i,t_1}\\ A_{i,t_1}g_{22,i,t_1}\\ -\mathbf{g}_{23,i,t_1}\end{bmatrix}\Big(V_{i,t_2}+U_{i,t_2}\Big) - \begin{bmatrix}K_{i,t_1}g_{2,i,t_1}-K_{i,t_2}\\ A_{i,t_1}g_{2,i,t_1}-A_{i,t_2}\\ -\mathbf{g}_{3,i,t_1}\end{bmatrix}g_{2,i,t_1},$$

and

$$\dot{\mathbf{m}}_2(\mathbf{w}_i,\nu_{1i}-z_{11i}\gamma,\mu_{2i},\gamma,\boldsymbol{\theta}) = \begin{bmatrix}K_{i,t_1}g_{12,i,t_1}\\ A_{i,t_1}g_{12,i,t_1}\\ -\mathbf{g}_{13,i,t_1}\end{bmatrix}\Big(V_{i,t_2}+U_{i,t_2}\Big) - \begin{bmatrix}K_{i,t_1}g_{2,i,t_1}-K_{i,t_2}\\ A_{i,t_1}g_{2,i,t_1}-A_{i,t_2}\\ -\mathbf{g}_{3,i,t_1}\end{bmatrix}g_{1,i,t_1}.$$

Then conditional on the corresponding covariates (and note that $\mathbf{z}_{1i}\supset\mathbf{z}_{2i}$),

$$\begin{aligned}\mathbb{E}[\dot{\mathbf{m}}_1(\mathbf{w}_i,\nu_{1i}-z_{11i}\gamma,\mu_{2i},\gamma,\boldsymbol{\theta})|\mathbf{z}_{1i}] &= -\begin{bmatrix}K_{i,t_1}g_{2,i,t_1}-K_{i,t_2}\\ A_{i,t_1}g_{2,i,t_1}-A_{i,t_2}\\ -\mathbf{g}_{3,i,t_1}\end{bmatrix}g_{2,i,t_1}\\ \mathbb{E}[\dot{\mathbf{m}}_2(\mathbf{w}_i,\nu_{1i}-z_{11i}\gamma,\mu_{2i},\gamma,\boldsymbol{\theta})|\mathbf{z}_{1i}] &= -\begin{bmatrix}K_{i,t_1}g_{2,i,t_1}-K_{i,t_2}\\ A_{i,t_1}g_{2,i,t_1}-A_{i,t_2}\\ -\mathbf{g}_{3,i,t_1}\end{bmatrix}g_{1,i,t_1}.\end{aligned}$$

Note that we can drop the conditional expectation since both K_{i,t_2} and A_{i,t_2} are determined by time-t information. One example would be $K_{i,t_2} = K_{i,t_1} + I_{i,t_1}$ if there is no depreciation, and $A_{i,t_2} = A_{i,t_1}+1$ if $t_2 - t_1 = 1$ and the unit of aging is calendar year.

Next we consider the two linear bias terms. First we consider the conditional correlation between $\dot{\mathbf{m}}_1(\mathbf{w}_i, \nu_{1i} - z_{11i}\gamma, \mu_{2i}, \gamma, \boldsymbol{\theta})$ and $\varepsilon_{1i} = U_{i,t_1}$. Since both K_{i,t_2} and A_{i,t_2} can be regarded as deterministic functions of time-t variables, it is true that

$$\begin{aligned}\mathbf{b}_{1,1,i} = \mathbb{E}\Big[\dot{\mathbf{m}}_1(\mathbf{w}_i, \nu_{1i} - z_{11i}\gamma, \mu_{2i}, \gamma, \boldsymbol{\theta})\varepsilon_{1i}\Big|\mathbf{z}_{1i}\Big] &= \begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1}\end{bmatrix} \mathbb{E}\Big[(V_{i,t_2} + U_{i,t_2})U_{i,t_1}\Big|(L,I,K,A)_{i,t_1}\Big] \\ &= \begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1}\end{bmatrix} \mathbb{C}\mathrm{ov}\Big[V_{i,t_2}, U_{i,t_1}\Big|(L,I,K,A)_{i,t_1}\Big].\end{aligned}$$

To prove the last line, note that $\mathbb{E}_{t_1}[U_{i,t_2}U_{i,t_1}] = 0$ by applying iterative expectation tp $U_{i,t}$. On the other hand, V_{i,t_2} may not be orthogonal to time-t_2 information, and hence it is generally impossible to conclude the last line is zero.

With the same logic, we have, for the other linear bias term,

$$\mathbf{b}_{1,2,i} = \mathbb{E}[\dot{\mathbf{m}}_2(\mathbf{w}_i, \nu_{1i} - z_{11i}\gamma, \mu_{2i}, \gamma, \boldsymbol{\theta})\varepsilon_{2i}|\mathbf{z}_{1i}] = \begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1}\end{bmatrix} \mathbb{C}\mathrm{ov}\Big[V_{i,t_2}, \chi_{i,t_2}\Big|(L,I,K,A)_{i,t_1}\Big].$$

And we note the above bias is generally not zero, since both V_{i,t_2} and χ_{i,t_2} depend on time-t_2 information.

For the quadratic bias term, we compute the second order derivatives

$$\begin{aligned}&\ddot{\mathbf{m}}_{11}(\mathbf{w}_i, \nu_{1i} - z_{11i}\gamma, \mu_{2i}, \gamma, \boldsymbol{\theta}) \\ &= \begin{bmatrix} K_{i,t_2} g_{222,i,t_1} \\ A_{i,t_2} g_{222,i,t_1} \\ -\mathbf{g}_{223,i,t_1}\end{bmatrix} \Big(V_{i,t_2} + U_{i,t_2}\Big) - 2\begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1}\end{bmatrix} g_{2,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1}\end{bmatrix} g_{22,i,t_1},\end{aligned}$$

$$\begin{aligned}&\ddot{\mathbf{m}}_{22}(\mathbf{w}_i, \nu_{1i} - z_{11i}\gamma, \mu_{2i}, \gamma, \boldsymbol{\theta}) \\ &= \begin{bmatrix} K_{i,t_1} g_{112,i,t_1} \\ A_{i,t_1} g_{112,i,t_1} \\ -\mathbf{g}_{113,i,t_1}\end{bmatrix} \Big(V_{i,t_2} + U_{i,t_2}\Big) - 2\begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1}\end{bmatrix} g_{1,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1}\end{bmatrix} g_{11,i,t_1}.\end{aligned}$$

$$\begin{aligned}&\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \nu_{1i} - z_{11i}\gamma, \mu_{2i}, \gamma, \boldsymbol{\theta}) \\ &= \begin{bmatrix} K_{i,t_1} g_{122,i,t_1} \\ A_{i,t_1} g_{122,i,t_1} \\ -\mathbf{g}_{123,i,t_1}\end{bmatrix} \Big(V_{i,t_2} + U_{i,t_2}\Big) - \begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1}\end{bmatrix} g_{1,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1}\end{bmatrix} g_{2,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1}\end{bmatrix} g_{12,i,t_1}\end{aligned}$$

Hence the three bias terms are

$$\mathbf{b}_{2,11,ij} = -\frac{1}{2}\left\{2\begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1}\end{bmatrix} g_{2,i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1}\end{bmatrix} g_{22,i,t_1}\right\} \mathbb{V}\Big[U_{i,t_1}\Big|(L,I,K,A)_{i,t_1}\Big],$$

$$\mathbf{b}_{2,22,ij} = -\frac{1}{2}\left\{2\begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1}\end{bmatrix} g_{1,i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1}\end{bmatrix} g_{11,i,t_1}\right\} \mathbb{V}\Big[\chi i, t_2\Big|(L,I,K,A)_{i,t_1}\Big],$$

$$\mathbf{b}_{2,12,ij} = -\left\{-\begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1}\end{bmatrix} g_{1,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1}\end{bmatrix} g_{2,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1}\end{bmatrix} g_{12,i,t_1}\right\} \mathbb{C}\mathrm{ov}\Big[U_{i,t_1}, \chi_{i,t_2}\Big|(L,I,K,A)_{i,t_1}\Big].$$

Again if $U_{i,t}$ is purely measurement error, the bias $\mathbf{b}_{2,12,ij}$ will be zero.

The last step is to recover the influence function. Since we use series expansion, it takes relatively simple form. The first piece, $\bar{\boldsymbol{\Psi}}_1$, comes from the moment condition, which is

$$\bar{\boldsymbol{\Psi}}_1 = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \mathbf{m}(\mathbf{w}_i, \mu_{1i}, \mu_{2i}, \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1}\end{bmatrix} \Big(V_{i,t_2} + U_{i,t_2}\Big).$$

The second piece, $\bar{\mathbf{\Psi}}_2$, can be decomposed into three, and two of them correspond to contributions of estimating ν_{1i} and μ_{2i} in the first step,

$$\bar{\mathbf{\Psi}}_{2,1} = -\frac{1}{\sqrt{n}}\mathbf{\Sigma}_0 \sum_i \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{2,i,t_1} U_{i,t_1},$$

$$\bar{\mathbf{\Psi}}_{2,2} = -\frac{1}{\sqrt{n}}\mathbf{\Sigma}_0 \sum_i \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{1,i,t_1} \left(\chi_{i,t_2} - P_{i,t_1}\right).$$

The final piece is the contribution of estimating β_L in the first step. For this purpose, we use results in Section SA-4.2. Then

$$\bar{\mathbf{\Psi}}_{2,3} = \frac{1}{\sqrt{n}} \frac{1}{\mathbb{E}\mathbb{V}[L_{i,t_1}|(I,K,A)_{i,t_1}]} \tilde{\mathbf{\Sigma}}_0 \sum_i \mathbb{E}\Big[L_{i,t_1}\Big|(I,K,A)_{i,t_1}\Big] U_{i,t_1}.$$

■

SA-9.16 Proposition SA.14

SA-9.16.1 Part 1

For the ease of exposition we ignore (asymptotic negligible) remainder terms in the proof. Then $\hat{\boldsymbol{\theta}}$ has the expansion

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \frac{1}{\sqrt{n}}\sum_i \mathbf{a}_i + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_i\left(\hat{\mu}_i - \mu_i\right) + \frac{1}{\sqrt{n}}\sum_i \mathbf{c}_i\left(\hat{\mu}_i - \mu_i\right)^2,$$

where to save notations we used

$$\begin{aligned}
\mathbf{a}_i &= -\left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \\
\mathbf{b}_i &= -\left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \\
\mathbf{c}_i &= -\frac{1}{2}\left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0).
\end{aligned}$$

Denote the leave-j-out estimator by $\hat{\boldsymbol{\theta}}^{(j)}$, it is easy to see that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}_0\right) = \frac{\sqrt{n}}{n-1}\sum_{i,i\neq j} \mathbf{a}_i + \frac{\sqrt{n}}{n-1}\sum_{i,i\neq j} \mathbf{b}_i\left(\hat{\mu}_i^{(j)} - \mu_i\right) + \frac{\sqrt{n}}{n-1}\sum_{i,i\neq j} \mathbf{c}_i\left(\hat{\mu}_i^{(j)} - \mu_i\right)^2.$$

Recall that the jackknife estimator is defined as

$$\hat{\boldsymbol{\theta}}^{(\cdot)} = \frac{1}{n}\sum_j \hat{\boldsymbol{\theta}}^{(j)},$$

and with some algebraic manipulation,

$$\begin{aligned}
(n-1)\cdot\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(\cdot)} - \hat{\boldsymbol{\theta}}\right) = {} & \frac{1}{\sqrt{n}}\sum_j \sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j - r_j\right) && \text{(I)} \\
& + \frac{1}{\sqrt{n}}\sum_j \sum_{i,i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(\hat{\mu}_j - r_j\right)^2 && \text{(II)} \\
& + \frac{2}{\sqrt{n}}\sum_j \sum_{i,i\neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i - \mu_i\right)\left(\hat{\mu}_j - r_j\right). && \text{(III)}
\end{aligned}$$

By Assumption A.2(2), we could ignore the approximation error. And (I) becomes

$$\text{(I)} = \frac{1}{\sqrt{n}}\sum_j \sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j - \mu_j + \mu_j - r_j\right)$$

$$= \underbrace{\frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_\ell \pi_{j\ell}\varepsilon_\ell \right)}_{\text{(I.1)}} \underbrace{- \frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}} \varepsilon_j}_{\text{(I.2)}} + o_{\mathbb{P}}(1).$$

Then we have the following conditional expectations:

$$\begin{aligned}
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{(I.1)}\right] &= \frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \frac{\pi_{ij}^2}{1-\pi_{jj}} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i \varepsilon_i\right] \\
&= -\frac{1}{\sqrt{n}} \left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0 \right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \left[\sum_i \mathbf{b}_{1,i} \pi_{ii} \right] + \frac{1}{\sqrt{n}} \sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i \varepsilon_i\right] \left(\sum_{j,j\neq i} \frac{\pi_{ij}^2}{1-\pi_{jj}} - \pi_{ii} \right) \\
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{(I.2)}\right] &= 0.
\end{aligned}$$

To further simplify, note that

$$\left| \frac{1}{\sqrt{n}} \sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i \varepsilon_i\right] \left(\sum_{j,j\neq i} \frac{\pi_{ij}^2}{1-\pi_{jj}} - \pi_{ii} \right) \right| \precsim \frac{1}{\sqrt{n}} \sum_i \left| \sum_{j,j\neq i} \frac{\pi_{ij}^2}{1-\pi_{jj}} - \pi_{ii} \right| \precsim \frac{1}{\sqrt{n}} \sum_i \pi_{ii}^2 = o_{\mathbb{P}}(1).$$

One could conduct variance calculation, which is tedious yet straightforward. Now we consider (II), which has the following expansion:

$$\text{(II)} = \frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \left(\hat{\mu}_j - \mu_j + \mu_j - r_j \right)^2 = \underbrace{\frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \left(\sum_{\ell,m} \pi_{j\ell}\pi_{jm}\varepsilon_\ell\varepsilon_m \right)}_{\text{(II.1)}}$$
$$+ \underbrace{\frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \varepsilon_j^2}_{\text{(II.2)}} \underbrace{- \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \left(\sum_\ell \pi_{j\ell}\varepsilon_\ell\varepsilon_j \right)}_{\text{(II.3)}} + o_{\mathbb{P}}(1).$$

Therefore

$$\left| \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{(II.1)}\right] \right| = \left| \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \sum_\ell \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \pi_{j\ell}^2 \right| \precsim_{\mathbb{P}} \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \sum_\ell \pi_{ij}^2 \pi_{j\ell}^2 \leq \frac{1}{\sqrt{n}} \sum_{j,\ell} \pi_{j\ell}^2 \pi_{jj} = o_{\mathbb{P}}(1),$$

and

$$\begin{aligned}
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{(II.2)}\right] &= \frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 + o_{\mathbb{P}}(1) \\
&= -\frac{1}{\sqrt{n}} \left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0 \right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \left[\sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2 \right] + \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \frac{\pi_{ij}^2 \pi_{jj}}{(1-\pi_{jj})^2} + o_{\mathbb{P}}(1) \\
&= -\frac{1}{\sqrt{n}} \left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0 \right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \left[\sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2 \right] + o_{\mathbb{P}}(1), \qquad \text{(E.36)}
\end{aligned}$$

and using (E.36) again,

$$\left| \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{(II.3)}\right] \right| = \left| \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \pi_{jj} \right| \precsim_{\mathbb{P}} \frac{1}{\sqrt{n}} \sum_{i,j} \pi_{ij}^2 \pi_{jj} = o_{\mathbb{P}}(1).$$

Finally (III) has the expansion:

$$\text{(III)} = \underbrace{\frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_{\ell,m} \pi_{i\ell}\pi_{jm}\varepsilon_\ell\varepsilon_m \right)}_{\text{III.1}} \underbrace{- \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_\ell \pi_{i\ell}\varepsilon_\ell\varepsilon_j \right)}_{\text{III.2}} + o_{\mathbb{P}}(1).$$

Again we consider the conditional expectations:

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.1}\right] = \frac{2}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\sum_{\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_\ell^2\right]\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}},$$
$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.2}\right] = -\frac{2}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_j^2\right]\frac{\pi_{ij}^2}{1-\pi_{jj}}.$$

Therefore using (E.36) and $\pi_{j'j'} \leq 1$

$$\begin{aligned}
\left|\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.1}\right] + \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.2}\right]\right| &= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_\ell^2\right]\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_\ell^2\right]\frac{\pi_{i\ell}^2}{1-\pi_{\ell\ell}}\right| + o_{\mathbb{P}}(1) \\
&= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_\ell^2\right]\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,j,\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_\ell^2\right]\pi_{ij}\pi_{i\ell}\pi_{j\ell}\right| + o_{\mathbb{P}}(1) \\
&= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_\ell^2\right]\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}}\right| + o_{\mathbb{P}}(1) \\
&\precsim_{\mathbb{P}} \frac{1}{\sqrt{n}}\sqrt{\sum_{i,\ell}\pi_{i\ell}^2}\sqrt{\sum_{i,\ell}\left(\sum_{j}\frac{\pi_{ij}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}}\right)^2} \\
&= \frac{\sqrt{k}}{\sqrt{n}}\sqrt{\sum_{i,\ell}\sum_{jj'}\frac{\pi_{ij}\pi_{j\ell}\pi_{jj}\pi_{ij'}\pi_{j'\ell}\pi_{j'j'}}{(1-\pi_{jj})(1-\pi_{j'j'})}} \precsim_{\mathbb{P}} \frac{\sqrt{k}}{\sqrt{n}}\sqrt{\sum_{jj'}\pi_{jj}\pi_{j'j'}\pi_{jj'}^2} = \frac{\sqrt{k}}{\sqrt{n}}\cdot o_{\mathbb{P}}(\sqrt{k}) = o_{\mathbb{P}}(1).
\end{aligned}$$

Therefore we showed the desired result. ■

SA-9.16.2 Part 2

First note that the jackknife variance estimator takes the form:

$$(n-1)\sum_{j}\left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)}\right)^2,$$

where for a (column) vector $\mathbf{v}$, we use $\mathbf{v}^2$ to denote $\mathbf{v}\mathbf{v}^\mathsf{T}$ to save space. Then the variance estimator could be rewritten as

$$\hat{\boldsymbol{\mathcal{V}}} = (n-1)\sum_{j}\left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}\right)^2 - \frac{1}{n-1}\left(\hat{\boldsymbol{\mathcal{B}}}\right)^2 = (n-1)\sum_{j}\left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}\right)^2 + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Next recall that

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}} = &\underbrace{\frac{1}{n-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}_{\text{(I)}} \underbrace{-\frac{1}{n-1}\mathbf{a}_j}_{\text{(II)}} \underbrace{-\frac{1}{n-1}\mathbf{b}_j\left(\hat{\mu}_j - \mu_j\right)}_{\text{(III)}} \underbrace{-\frac{1}{n-1}\mathbf{c}_j\left(\hat{\mu}_j - \mu_j\right)^2}_{\text{(IV)}} \\
&\underbrace{+\frac{1}{n-1}\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j - r_j\right)}_{\text{(V)}} \underbrace{+\frac{1}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j - r_j\right)^2}_{\text{(VI)}} \underbrace{+\frac{2}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i - \mu_i\right)\left(\hat{\mu}_j - r_j\right)}_{\text{(VII)}}.
\end{aligned}$$

Therefore we have to consider the square of each term, as well as their interactions. As the proof is quite tedious, we list the main steps here. First we would like to recover the variance terms in Theorem SA.5 with

$$(n-1)\sum_{j}\left(\text{II}\right)^2 = \mathbb{V}[\bar{\boldsymbol{\Psi}}_1] + o_{\mathbb{P}}(1),\ (n-1)\sum_{j}\left(\text{II}\right)\left(\text{V}\right)^\mathsf{T} = \mathbb{C}\text{ov}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_1, \bar{\boldsymbol{\Psi}}_2] + o_{\mathbb{P}}(1),\ (n-1)\sum_{j}\left(\text{V}\right)^2 = \mathbb{V}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_2] + o_{\mathbb{P}}(1).$$

Furthermore, all the other square terms and interactions are asymptotically negligible. We use the following fact repeatedly: For two sequences $\{u_i\}$ and $\{v_j\}$,

$$\left|\sum_{i,j} u_i \pi_{ij} v_j\right| \leq \sqrt{\sum_i u_i^2}\sqrt{\sum_i \left(\sum_j \pi_{ij} v_j\right)^2} \leq \sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2}.$$

Term (I):

$$(n-1)\sum_j (\mathrm{I})^2 = \frac{1}{n-1}\sum_j \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^2 \asymp \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^2 = o_{\mathbb{P}}(1),$$

by consistency. Then it is also easy to show that for $\dagger = \mathrm{II}, \cdots, \mathrm{VII}$

$$(n-1)\sum_j (\mathrm{I})(\dagger)^{\mathsf{T}} = (\mathrm{I})\frac{1}{n-1}\sum_j (\dagger)^{\mathsf{T}} = o_{\mathbb{P}}(1)\cdot\frac{1}{n-1}\sum_j (\dagger)^{\mathsf{T}} = o_{\mathbb{P}}(1),$$

since the summands are bounded in probability uniformly in j.

Next term (II):

$$(n-1)\sum_j (\mathrm{II})^2 = \frac{1}{n-1}\sum_j \mathbf{a}_j^2,$$

which is asymptotically equivalent to $\mathbb{V}[\bar{\boldsymbol{\Psi}}_1]$ in Theorem SA.5. Now we consider the interactions:

$$\left|(n-1)\sum_j (\mathrm{II})(\mathrm{III})^{\mathsf{T}}\right| = \left|\frac{1}{n-1}\sum_j \mathbf{a}_j \mathbf{b}_j^{\mathsf{T}}\left(\hat{\mu}_j - \mu_j\right)\right| \leq o_{\mathbb{P}}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{a}_j \mathbf{b}_j^{\mathsf{T}}| = o_{\mathbb{P}}(1).$$

Similar techniques can be used to establish the following

$$(n-1)\sum_j (\mathrm{II})(\mathrm{IV})^{\mathsf{T}} = o_{\mathbb{P}}(1).$$

The interactions between (II) and (V), (VI) and (VII) are more involved. We first consider the interaction between (II) and (V):

$$\begin{aligned}
(n-1)\sum_j (\mathrm{II})(\mathrm{V})^{\mathsf{T}} &= -\frac{1}{n}\sum_j \mathbf{a}_j \varepsilon_j \sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}} + o_{\mathbb{P}}(1) && \text{(Assumption A.2(1))}\\
&= \frac{1}{n}\sum_j \mathbf{a}_j \varepsilon_j \sum_{i,i\neq j} \mathbf{b}_i \pi_{ij} - \frac{1}{n}\sum_j \mathbf{a}_j \varepsilon_j \sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} + o_{\mathbb{P}}(1)\\
&= \frac{1}{n}\sum_j \mathbf{a}_j \varepsilon_j \sum_{i,i\neq j} \mathbf{b}_i \pi_{ij} + o_{\mathbb{P}}(1),
\end{aligned}$$

which is asymptotically equivalent to $\mathbb{C}\mathrm{ov}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_1, \bar{\boldsymbol{\Psi}}_2]$. And by symmetry, $(n-1)\sum_j (\mathrm{V})(\mathrm{II})^{\mathsf{T}}$ is equivalent to $\mathbb{C}\mathrm{ov}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_2, \bar{\boldsymbol{\Psi}}_1]$. And as a short digression,

$$\begin{aligned}
(n-1)\sum_j (\mathrm{V})^2 &= \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + o_{\mathbb{P}}(1)\\
&= \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i] \frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \left(\mathbf{b}_i - \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\right) \frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\\
&\quad + \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i] \frac{\pi_{ij}}{1-\pi_{jj}}\right)\left(\sum_{i,i\neq j} \left(\mathbf{b}_i - \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\right) \frac{\pi_{ij}}{1-\pi_{jj}}\right)^{\mathsf{T}} + o_{\mathbb{P}}(1),
\end{aligned}$$

where the first term in the above display recovers $\mathbb{V}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_2]$, while the rest two are negligible by conditional expectation calculation. Therefore we recovered the asymptotic variance.

Back to the interaction terms,

$$\Big|(n-1)\sum_j (\mathrm{II})(\mathrm{VI})^\mathsf{T}\Big| = \Big|\frac{1}{n-1}\sum_j \mathbf{a}_j \sum_{i,i\neq j} \mathbf{c}_i^\mathsf{T}\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\Big(\hat\mu_j - r_j\Big)^2\Big| \precsim_\mathbb{P} \frac{1}{n-1}\sum_{i,j}\pi_{ij}^2 = o_\mathbb{P}(1),$$

and

$$\begin{aligned}
&\Big|(n-1)\sum_j (\mathrm{II})(\mathrm{VII})^\mathsf{T}\Big| = \Big|\frac{2}{n-1}\sum_j \mathbf{a}_j\Big(\hat\mu_j - r_j\Big)\sum_{i,i\neq j}\mathbf{c}_i^\mathsf{T}\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat\mu_i - \mu_i\Big)\Big| \\
&\precsim_\mathbb{P} \Big|\frac{2}{n-1}\sum_{i,j}\mathbf{a}_j\Big(\hat\mu_j - r_j\Big)\mathbf{c}_i^\mathsf{T}\pi_{ij}\Big(\hat\mu_i - \mu_i\Big)\Big| \qquad \text{(Assumption A.3(2))} \\
&\le \frac{2}{n-1}\cdot\sqrt{\sum_j |\mathbf{a}_j|^2\Big(\hat\mu_j - r_j\Big)^2}\sqrt{\sum_j |\mathbf{c}_j|^2\Big(\hat\mu_j - \mu_j\Big)^2} \le o_\mathbb{P}(1)\cdot\frac{2}{n-1}\cdot\sqrt{\sum_j |\mathbf{a}_j|^2\Big(\hat\mu_j - r_j\Big)^2}\sqrt{\sum_j |\mathbf{c}_j|^2} = o_\mathbb{P}(1),
\end{aligned}$$

With a quick inspection, the above method also applies to the following interactions

$$\begin{aligned}
&(n-1)\sum_j (\mathrm{III})(\mathrm{V})^\mathsf{T} = o_\mathbb{P}(1), \quad (n-1)\sum_j (\mathrm{III})(\mathrm{VI})^\mathsf{T} = o_\mathbb{P}(1), \quad (n-1)\sum_j (\mathrm{III})(\mathrm{VII})^\mathsf{T} = o_\mathbb{P}(1), \\
&(n-1)\sum_j (\mathrm{IV})(\mathrm{V})^\mathsf{T} = o_\mathbb{P}(1), \quad (n-1)\sum_j (\mathrm{IV})(\mathrm{VI})^\mathsf{T} = o_\mathbb{P}(1), \quad (n-1)\sum_j (\mathrm{IV})(\mathrm{VII})^\mathsf{T} = o_\mathbb{P}(1).
\end{aligned}$$

Next we consider the squared terms involving (III) and (IV):

$$\begin{aligned}
(n-1)\sum_j (\mathrm{III})^2 &= \frac{1}{n-1}\sum_j (\mathbf{b}_j)^2\Big(\hat\mu_j - \mu_j\Big)^2 \le o_\mathbb{P}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{b}_j|^2 = o_\mathbb{P}(1), \\
(n-1)\sum_j (\mathrm{IV})^2 &= \frac{1}{n-1}\sum_j (\mathbf{c}_j)^2\Big(\hat\mu_j - \mu_j\Big)^4 \le o_\mathbb{P}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{c}_j|^2 = o_\mathbb{P}(1).
\end{aligned}$$

What remains are $(\mathrm{V})(\mathrm{VI})^\mathsf{T}$, $(\mathrm{V})(\mathrm{VII})^\mathsf{T}$, $(\mathrm{VI})^2$, $(\mathrm{VI})(\mathrm{VII})^\mathsf{T}$ and $(\mathrm{VII})^2$.

$$\Big|(n-1)\sum_j (\mathrm{V})(\mathrm{VI})^\mathsf{T}\Big| = \Big|\frac{1}{n-1}\sum_{i,j}\mathbf{b}_i\pi_{ij}\Big(\hat\mu_j - r_j\Big)^3\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\left(\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\right)^2\right)^\mathsf{T}\Big| + o_\mathbb{P}(1) \precsim_\mathbb{P} \sqrt{\frac{1}{n}\sum_{j,i,\ell}\pi_{ij}^2\pi_{\ell j}^2} = o_\mathbb{P}(1).$$

And

$$\begin{aligned}
\Big|(n-1)\sum_j (\mathrm{V})(\mathrm{VII})^\mathsf{T}\Big| &= \Big|\frac{2}{n-1}\sum_j\left(\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat\mu_j - r_j\Big)\right)\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\Big(\hat\mu_\ell - \mu_\ell\Big)\Big(\hat\mu_j - r_j\Big)\right)^\mathsf{T}\Big| \\
&\precsim_\mathbb{P} \sqrt{\frac{1}{n-1}\sum_j\Big(\hat\mu_j - r_j\Big)^4\left|\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\Big(\hat\mu_\ell - \mu_\ell\Big)\right|^2} = o_\mathbb{P}(1),
\end{aligned}$$

where the last line uses Assumption A.2(1). Using techniques in the above results, we can show

$$(n-1)\sum_j (\mathrm{VI})^2 = o_\mathbb{P}(1), \quad (n-1)\sum_j (\mathrm{VII})^2 = o_\mathbb{P}(1), \quad (n-1)\sum_j (\mathrm{VI})(\mathrm{VII})^\mathsf{T} = o_\mathbb{P}(1),$$

which closes the proof. ■

SA-9.17 Lemma SA.15

Recall that $\hat\mu_i^\star = \hat\mu_i + \sum_j \pi_{ij}e_j^\star(r_j - \hat\mu_j)$, and $\mathbb{P}^\star[\cdot]$ denotes probability conditional on the original data. Since the bootstrap weights have zero mean and sub-Gaussian tail, the Hoeffding's inequality (Vershynin, 2018, Theorem 2.6.2), conditional on the original data, implies

$$\mathbb{P}^\star\left[\max_{1\le i\le n}|\hat\mu_i^\star - \hat\mu_i| \ge t\right] = \mathbb{P}^\star\left[\max_{1\le i\le n}\left|\sum_j \pi_{ij}e_j^\star(r_j - \hat\mu_j)\right| \ge t\right] \le n\cdot\max_{1\le i\le n}\mathbb{P}^\star\left[\left|\sum_j \pi_{ij}e_j^\star(r_j - \hat\mu_j)\right| \ge t\right]$$

$$
\begin{aligned}
&\leq n \cdot \max_{1\leq i\leq n} 2\exp\left(-\frac{Ct^2}{M^2\sum_j \pi_{ij}^2 (r_j-\hat{\mu}_j)^2}\right)\\
&\leq n \cdot \max_{1\leq i\leq n} 2\exp\left(-\frac{Ct^2}{M^2(\sum_j \pi_{ij}^2)(\max_{1\leq i\leq n}(r_i-\hat{\mu}_i)^2)}\right)\\
&= n \cdot \max_{1\leq i\leq n} 2\exp\left(-\frac{Ct^2}{M^2\pi_{ii}(\max_{1\leq i\leq n}(r_i-\hat{\mu}_i)^2)}\right)\\
&\leq 2\exp\left(-\frac{Ct^2}{M^2(\max_{1\leq i\leq n}\pi_{ii})(\max_{1\leq i\leq n}(r_i-\hat{\mu}_i)^2)}+\log(n)\right),
\end{aligned}
$$

where $M = \inf\{t \geq 0 : \ \mathbb{E}[\exp(e_i^{\star 2}/t^2)] \leq 2\} < \infty$, and we also used the fact that $\sum_j \pi_{ij}^2 = \pi_{ii}$. Therefore, for all $t > 0$, $\mathbb{P}^\star[\max_{1\leq i\leq n} |\hat{\mu}_i^\star - \hat{\mu}_i| \geq t] = o_{\mathbb{P}}(1)$ provided that

$$
\Big(\max_{1\leq i\leq n} \pi_{ii}\Big)\Big(\max_{1\leq i\leq n} (r_i-\hat{\mu}_i)^2\Big) = o_{\mathbb{P}}\Big(\frac{1}{\log n}\Big).
$$

Since the conditional probability is always bounded by one, the unconditional probability converges to zero by dominated convergence under the above condition. Finally, note that

$$
\max_{1\leq i\leq n} (r_i-\hat{\mu}_i)^2 \leq 2\max_{1\leq i\leq n} \varepsilon_i^2 + 2\max_{1\leq i\leq n} (\hat{\mu}_i-\mu_i)^2,
$$

so that the desired result follows from the assumptions of the lemma. ■

SA-9.18 Proposition SA.16

First we note that consistency can be established with the same argument used for Theorem SA.1, hence is omitted here. Given consistency, we are able to linearize the estimating equation (E.33) with respect to $\hat{\boldsymbol{\theta}}^\star$, around $\hat{\boldsymbol{\theta}}$:

$$
\sqrt{n}\Big(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}}\Big) = \boldsymbol{\Sigma}_0 \left[\frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i, \hat{\mu}_i^\star, \hat{\boldsymbol{\theta}})\right]\Big(1+o_{\mathbb{P}}(1)\Big),
$$

where for notational simplicity, we define $\mathbf{m}^\star(\mathbf{w}_i,\cdot,\cdot) := (1+e_i^\star)\cdot\mathbf{m}(\mathbf{w}_i,\cdot,\cdot)$. We further expand the above with respect to the bootstrapped first step:

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i, \hat{\mu}_i^\star, \hat{\boldsymbol{\theta}}) &= \frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) && \text{(E.41)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big) && \text{(E.42)}\\
&+ \frac{1}{\sqrt{n}}\sum_i \frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)^2 + o_{\mathbb{P}}(1). && \text{(E.43)}
\end{aligned}
$$

Analyses of the above terms are similar to those of Lemma SA.3 and SA.4, with more delicate arguments.

SA-9.18.1 Term (E.41)

Lemma SA.22 (Term (E.41)).
Assume A.1, A.2 and A.4 hold, and $k = O(\sqrt{n})$. Then

$$
\text{(E.41)} = \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathbb{P}}(1).
$$

⌟

Note that

$$
\text{(E.41)} = \frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) + o_{\mathbb{P}}(1).
$$

The second equality uses (E.5), while the last one comes from the argument that

$$\frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big) \precsim_{\mathbb{P}} \frac{1}{n}\sum_i e_i^\star \cdot \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}}) \to_{\mathbb{P}} \mathbb{E}\left[e_i^\star \cdot \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\right],$$

given Assumption A.1(5). To further understand the last term, we still need to expand it with respect to $\hat{\mu}_i$, yielding

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) && \text{(I)}\\ &+ \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_i-\mu_i\Big) && \text{(II)}\\ &+ \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_i-\mu_i\Big)^2\cdot(1+o_{\mathbb{P}}(1)). && \text{(III)}\end{aligned}$$

(I) apparently contributes to the first order. For (II), note that it can be simplified using exactly the same argument used in Lemma SA.3 and SA.3. Equivalently, assuming A.1 and A.2, then

$$\text{(II)} = O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathbb{P}}(1).$$

By the same argument, (III) can be simplified with Lemma SA.4 and SA.4. Namely, assume A.1 and A.2 hold, then

$$\text{(III)} = O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathbb{P}}(1).$$

■

SA-9.18.2 Term (E.42)

Lemma SA.23 (Term (E.42)).
Assume A.1, A.2 and A.4 hold, and $k = O(\sqrt{n})$. Then

$$\text{(E.42)} = \frac{1}{\sqrt{n}}\sum_i\left(\sum_j \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j,\mu_j,\boldsymbol{\theta}_0)|\mathbf{z}_j\right]\pi_{ij}\right)\varepsilon_i e_i^\star + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{1,i}\cdot\pi_{ii} + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_{1,i}$ is given in Lemma SA.3. ⌟

For (E.42), we first show that it is possible to replace $\hat{\boldsymbol{\theta}}$ by $\boldsymbol{\theta}_0$, provided $\partial\dot{\mathbf{m}}/\partial\boldsymbol{\theta}$ is Hölder continuous in μ_i and $\boldsymbol{\theta}$:

$$\text{(E.42)} = \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_i^\star-\hat{\mu}_i\Big) + \frac{1}{n}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star-\hat{\mu}_i\Big)\sqrt{n}\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big),$$

where the second term is bounded by the following

$$\left|\frac{1}{n}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star-\hat{\mu}_i\Big)\sqrt{n}\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big)\right| \precsim_{\mathbb{P}} \frac{1}{n}\sum_i\left|\frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star-\hat{\mu}_i\Big)\right| = o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_i\left|\frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\right| = o_{\mathbb{P}}(1),$$

where the last one uses the uniform consistency of $\hat{\mu}_i^\star$ and $\hat{\mu}_i$. Hence

$$\begin{aligned}\text{(E.42)} &= \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_i^\star-\hat{\mu}_i\Big) + o_{\mathbb{P}}(1)\\ &= \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\Big(\sum_j \pi_{ij}\varepsilon_j e_j^\star\Big) \underbrace{-\frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\Big(\sum_j \pi_{ij}(\hat{\mu}_j-\mu_j)e_j^\star\Big)}_{\text{(I)}} + o_{\mathbb{P}}(1).\end{aligned}$$

For (I),

$$\mathbb{E}^\star\left[(\text{I})(\text{I})^\intercal\right] = \frac{1}{n}\mathbb{E}^\star\left[\sum_{i,i',j,j'}\dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\dot{\mathbf{m}}^\star(\mathbf{w}_{i'},\hat{\mu}_{i'},\boldsymbol{\theta}_0)^\intercal(\hat{\mu}_j-\mu_j)(\hat{\mu}_{j'}-\mu_{j'})e_j^\star e_{j'}^\star\pi_{ij}\pi_{i'j'}\right]$$

$$= \frac{1}{n} \sum_{\substack{i,i',j \\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \dot{\mathbf{m}}(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)^\mathsf{T} (\hat{\mu}_j - \mu_j)^2 \pi_{ij} \pi_{i'j} \tag{II}$$

$$+ \frac{2}{n} \sum_{\substack{i,i' \\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \dot{\mathbf{m}}(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)^\mathsf{T} (\hat{\mu}_i - \mu_i)(\hat{\mu}_{i'} - \mu_{i'}) \pi_{ii} \pi_{i'i'} \tag{III}$$

$$+ \frac{2}{n} \sum_{\substack{i,j \\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)^\mathsf{T} (\hat{\mu}_j - \mu_j)^2 \pi_{ij}^2 \tag{IV}$$

$$+ \frac{C_1}{n} \sum_{\substack{i,j \\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \dot{\mathbf{m}}(\mathbf{w}_j, \hat{\mu}_j, \boldsymbol{\theta}_0)^\mathsf{T} (\hat{\mu}_j - \mu_j)^2 \pi_{ij} \pi_{jj} \tag{V}$$

$$+ \frac{C_2}{n} \sum_{\substack{i \\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)^\mathsf{T} (\hat{\mu}_i - \mu_i)^2 \pi_{ii}^2, \tag{VI}$$

where C_1 and C_2 are related to the third and fourth moments of $e_i^\star$. Then for each term,

$$\begin{aligned} |\text{(II)}| &\le \left(\max_{1 \le i \le n} |\hat{\mu}_i - \mu_i|^2 \right) \cdot \frac{1}{n} \sum_{\substack{i,i' \\ \text{distinct}}} |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)| \, |\dot{\mathbf{m}}(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)| \, \pi_{ii'} \\ &\le o_\mathbb{P}(1) \cdot \frac{1}{n} \sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 = o_\mathbb{P}(1), \qquad \text{(projection and Assumption A.2(1))} \end{aligned}$$

provided $\dot{\mathbf{m}}$ is Hölder continuous in μ_i. (III) can be handled by observing that

$$|\text{(III)}| \le \left(\frac{1}{\sqrt{n}} \sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)| \, \pi_{ii} \, |\hat{\mu}_i - \mu_i| \right)^2 \le o_\mathbb{P}(1) \cdot \left(\frac{1}{\sqrt{n}} \sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)| \, \pi_{ii} \right)^2 = o_\mathbb{P}\left(\frac{k^2}{n} \right).$$

Similarly

$$|\text{(IV)}| \le o_\mathbb{P}(1) \cdot \frac{2}{n} \sum_{\substack{i,j \\ \text{distinct}}} |\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|^2 \, \pi_{ij}^2 = o_\mathbb{P}\left(\frac{k}{n} \right),$$

and

$$\begin{aligned} |\text{(V)}| &\le \frac{C_1}{n} \left(\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 \right)^{1/2} \left(\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 \, |\hat{\mu}_i - \mu_i|^4 \pi_{jj}^2 \right)^{1/2} \\ &\precsim_\mathbb{P} n^{-1} \cdot \sqrt{n} \cdot \sqrt{k} \cdot o_\mathbb{P}(1) = o_\mathbb{P}\left(\sqrt{\frac{k}{n}} \right). \end{aligned}$$

Finally,

$$|\text{(VI)}| \le \frac{C_2}{n} \sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 \, |\hat{\mu}_i - \mu_i|^2 \pi_{ii}^2 = o_\mathbb{P}\left(\frac{k}{n} \right).$$

To summarize, we have the following

$$\begin{aligned} \text{(E.42)} &= \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_j e_j^\star \Big) + o_\mathbb{P}\left(\frac{k}{\sqrt{n}} \vee 1 \right) \\ &= \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_j e_j^\star \Big) + o_\mathbb{P}\left(\frac{k}{\sqrt{n}} \vee 1 \right), \end{aligned}$$

where the second line relies on almost the same argument. Finally, we can apply the same techniques used to prove Lemma SA.3 and SA.3, yielding

$$\text{(E.42)} = \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0) | \mathbf{z}_j \right] \pi_{ij} \right) \varepsilon_i e_i^\star + \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + o_\mathbb{P}\left(\frac{k}{\sqrt{n}} \vee 1 \right).$$

■

SA-9.18.3 Term (E.43)

Lemma SA.24 (Term (E.43)).
Assume A.1, A.2 and A.4 hold, and $k = O(\sqrt{n})$. Then

$$(\text{E.43}) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + \frac{1}{\sqrt{n}} \sum_{i} \mathbf{b}_{2,ii} \cdot \pi_{ii}^2 \cdot \mathbb{E}[e_i^{\star 3}] + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_{2,ij}$ is given in Lemma SA.4. ⌟

First note that

$$\begin{aligned}(\text{E.43}) &= \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\ &= \underbrace{\frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2}_{\text{(I)}} + \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \left[\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) - \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\right] \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2,\end{aligned}$$

where the second term is easily bounded by

$$\begin{aligned}&\left| \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \left[\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) - \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\right] \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \right| \\ &\leq \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} (1 + e_i) \cdot \mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}}) \cdot (|\tilde{\mu}_i^\star - \mu_i| + |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|)^\alpha \cdot |\hat{\mu}_i^\star - \hat{\mu}_i|^2 \\ &\leq o_{\mathbb{P}}(1) \cdot \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} (1 + e_i) \cdot \mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}}) \cdot |\hat{\mu}_i^\star - \hat{\mu}_i|^2. \qquad \text{(II)}\end{aligned}$$

Compare (I) and (II) and note that Assumption A.1(7) imposes the same restrictions on $\ddot{\mathbf{m}}$ and $\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}})$. Hence generically, (II) has the order

$$(\text{II}) = o_{\mathbb{P}}\left(|(\text{I})|\right).$$

Next we consider (I), which can be written as

$$(\text{I}) = \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \hat{\varepsilon}_j e_j^\star \Big)^2 = \frac{1}{\sqrt{n}} \sum_{i,j,\ell} \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \hat{\varepsilon}_j \hat{\varepsilon}_\ell e_j^\star e_\ell^\star \pi_{ij} \pi_{i\ell}.$$

The key step, as before, is to replace $\hat{\varepsilon}$ by ε. Note that

$$\begin{aligned}(\text{I}) = &\frac{1}{\sqrt{n}} \sum_{i,j,\ell} \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \hat{\varepsilon}_j \varepsilon_\ell e_j^\star e_\ell^\star \pi_{ij} \pi_{i\ell} \\ &- \frac{1}{\sqrt{n}} \sum_{i,\ell} \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) (\hat{\mu}_i^\star - \hat{\mu}_i)(\hat{\mu}_\ell - \mu_\ell) e_\ell^\star \pi_{i\ell}, \qquad \text{(III)}\end{aligned}$$

and (for simplicity let $\mathbf{a}_i^\star = \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(\hat{\mu}_i^\star - \hat{\mu}_i)$)

$$\begin{aligned}\mathbb{E}^\star\left[(\text{III})(\text{III})^\mathsf{T}\right] &= \frac{1}{4n} \sum_{i,i',j,j'} \mathbb{E}^\star\left[\mathbf{a}_i^\star \mathbf{a}_{i'}^{\star\mathsf{T}} (\hat{\mu}_j - \mu_j)(\hat{\mu}_{j'} - \mu_{j'}) e_j^\star e_{j'}^\star \pi_{ij} \pi_{i'j'}\right] \\ &= \frac{1}{4n} \sum_{\substack{i,i',j \\ \text{distinct}}} \mathbb{E}^\star\left[\mathbf{a}_i^\star \mathbf{a}_{i'}^{\star\mathsf{T}}\right] (\hat{\mu}_j - \mu_j)^2 \pi_{ij} \pi_{i'j} \qquad \text{(IV)} \\ &\quad + \frac{1}{4n} \sum_{\substack{i,j \\ \text{distinct}}} \mathbb{E}^\star\left[\mathbf{a}_i^\star \mathbf{a}_i^{\star\mathsf{T}}\right] (\hat{\mu}_j - \mu_j)^2 \pi_{ij}^2 \qquad \text{(V)} \\ &\quad + \frac{1}{2n} \sum_{\substack{i,i' \\ \text{distinct}}} \mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star] \mathbb{E}^\star[\mathbf{a}_{i'}^\star e_{i'}^\star]^\mathsf{T} (\hat{\mu}_i - \mu_i)(\hat{\mu}_{i'} - \mu_{i'}) \pi_{ii} \pi_{i'i'} \qquad \text{(VI)}\end{aligned}$$

$$+\frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]\mathbb{E}^\star\left[e_{i'}^{\star 2}\mathbf{a}_{i'}^{\star\mathsf{T}}\right](\hat{\mu}_{i'}-\mu_{i'})^2\pi_{ii'}\pi_{i'i'} \tag{VII}$$

$$+\frac{1}{4n}\sum_i\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_i^{\star\mathsf{T}}e_i^{\star 2}\right](\hat{\mu}_i-\mu_i)^2\pi_{ii}^2. \tag{VIII}$$

Then

$$|(\mathrm{IV})| = \left|\frac{1}{4n}\sum_{\substack{i,i',j\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathsf{T}}\right](\hat{\mu}_j-\mu_j)^2\pi_{ij}\pi_{i'j}\right| \precsim o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_{i,i'}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathsf{T}}\right]\pi_{ii'} \le o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_{i,i'}\mathbb{E}^\star\left[|\mathbf{a}_i^\star|\right]\mathbb{E}^\star\left[|\mathbf{a}_{i'}^\star|\right]\pi_{ii'}$$
$$\le o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_{i,i'}|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)||\ddot{\mathbf{m}}(\mathbf{w}_{i'},\mu_{i'},\boldsymbol{\theta}_0)|\pi_{ii'} \le o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2 = o_{\mathbb{P}}(1),$$

where the second line uses Assumption A.2(1), the fourth line uses Assumption A.4(2), and the last line uses projection property and Assumption A.1(7). Similarly, we have, for (V),

$$|(\mathrm{V})| = \left|\frac{1}{4n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_i^{\star\mathsf{T}}\right](\hat{\mu}_j-\mu_j)^2\pi_{ij}^2\right| \precsim o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_{i,j}|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\pi_{ij}^2 = o_{\mathbb{P}}\left(\frac{k}{n}\right),$$

and the last equality is a simple consequence of Assumption A.1(7). (VI) is the most difficult, which can be rewritten as

$$|(\mathrm{VI})| = \frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star]\mathbb{E}^\star[\mathbf{a}_{i'}^\star e_{i'}^\star]^{\mathsf{T}}(\hat{\mu}_i-\mu_i)(\hat{\mu}_{i'}-\mu_{i'})\pi_{ii}\pi_{i'i'}$$
$$\asymp\left(\frac{1}{\sqrt{n}}\sum_i\mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star](\hat{\mu}_i-\mu_i)\pi_{ii}\right)^2 \precsim o_{\mathbb{P}}(1)\cdot\left(\frac{1}{\sqrt{n}}\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\pi_{ii}\right)^2 = o_{\mathbb{P}}\left(\frac{k^2}{n}\right).$$

And

$$|(\mathrm{VII})| = \left|\frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]\mathbb{E}^\star\left[e_{i'}^{\star 2}\mathbf{a}_{i'}^{\star\mathsf{T}}\right](\hat{\mu}_{i'}-\mu_{i'})^2\pi_{ii'}\pi_{i'i'}\right|$$
$$\precsim\frac{1}{n}\left(\sum_i|\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]|^2\right)^{1/2}\left(\sum_i|\mathbb{E}^\star\left[e_i^{\star 2}\mathbf{a}_i^\star\right]|^2|\hat{\mu}_i-\mu_i|^2\pi_{ii}^2\right)^{1/2} \qquad \text{(projection)}$$
$$\le o_{\mathbb{P}}(1)\cdot\frac{1}{n}\left(\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\right)^{1/2}\left(\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\pi_{ii}^2\right)^{1/2} = o_{\mathbb{P}}(1)\cdot n^{-1}\cdot n^{1/2}\cdot k^{1/2} = o_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right).$$

Finally

$$|(\mathrm{VII})| = \frac{1}{4n}\sum_i\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_i^{\star\mathsf{T}}e_i^{\star 2}\right](\hat{\mu}_i-\mu_i)^2\pi_{ii}^2 \precsim o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\pi_{ii}^2 = o_{\mathbb{P}}\left(\frac{k}{n}\right).$$

Hence we have shown that

$$(\mathrm{I}) = \frac{1}{\sqrt{n}}\sum_{i,j,\ell}\frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\hat{\varepsilon}_j\varepsilon_\ell e_j^\star e_\ell^\star\pi_{ij}\pi_{i\ell}+o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}}\vee 1\right).$$

Not surprisingly, we can replicate the above argument, and replace $\hat{\varepsilon}_j$ by ε_j in the above display, yielding

$$(\mathrm{I}) = \frac{1}{\sqrt{n}}\sum_{i,j,\ell}\frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_j\varepsilon_\ell e_j^\star e_\ell^\star\pi_{ij}\pi_{i\ell}+o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}}\vee 1\right).$$

The next step is to apply Lemma SA.4 to conclude that

$$\begin{aligned}(\mathrm{I}) &= \frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}^{\star}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_j^2 e_j^{\star 2}\right]\pi_{ij}^2 + o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}}\vee 1\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{b}_{2,ij}\cdot\pi_{ij}^2 + \frac{1}{\sqrt{n}}\sum_{i}\mathbf{b}_{2,ii}\cdot\pi_{ii}^2\cdot\mathbb{E}[e_i^{\star 3}] + o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}}\vee 1\right).\end{aligned}$$

■

SA-9.18.4 Asymptotic Representation

This is a simple consequence of linearization, Lemma SA.22, SA.23 and SA.24. ■

SA-9.19 Proposition SA.17

SA-9.19.1 Part 1

For the ease of exposition we ignore (asymptotic negligible) remainder terms in the proof. Then $\hat{\boldsymbol{\theta}}^{\star}$ has the expansion

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star}-\hat{\boldsymbol{\theta}}\right) = \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^{\star}\hat{\mathbf{a}}_i + \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^{\star}\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star}-\hat{\mu}_i\right) + \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^{\star}\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star}-\hat{\mu}_i\right)^2,$$

where to save notations we used $\omega_i^{\star}=1+e_i^{\star}$, $n_\omega=\sum_i\omega_i^{\star}$, and

$$\hat{\mathbf{a}}_i=\boldsymbol{\Sigma}_0\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{b}}_i=\boldsymbol{\Sigma}_0\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{c}}_i=\boldsymbol{\Sigma}_0\frac{\ddot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})}{2}.$$

For future reference, let

$$\mathbf{a}_i=\boldsymbol{\Sigma}_0\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) \qquad \mathbf{b}_i=\boldsymbol{\Sigma}_0\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) \qquad \mathbf{c}_i=\boldsymbol{\Sigma}_0\frac{\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)}{2}.$$

Denote the leave-j-out estimator by $\hat{\boldsymbol{\theta}}^{\star,(j)}$, it is easy to see that

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(j)}-\hat{\boldsymbol{\theta}}\right) &= \frac{\sqrt{n}}{n_\omega-1}\sum_i(\omega_i^{\star}-\delta_{ij})\hat{\mathbf{a}}_i + \frac{\sqrt{n}}{n_\omega-1}\sum_i(\omega_i^{\star}-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)\\ &\quad + \frac{\sqrt{n}}{n_\omega-1}\sum_i(\omega_i^{\star}-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)^2,\end{aligned}$$

where $\delta_{ij}=\mathbb{1}[i=j]$. Recall that the jackknife estimator is defined as

$$\hat{\boldsymbol{\theta}}^{\star,(\cdot)}=\frac{1}{n_\omega}\sum_j\omega_j^{\star}\hat{\boldsymbol{\theta}}^{\star,(j)},$$

hence

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)}-\hat{\boldsymbol{\theta}}\right) &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^{\star}(\omega_i^{\star}-\delta_{ij})\hat{\mathbf{a}}_i + \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^{\star}(\omega_i^{\star}-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)\\ &\quad + \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^{\star}(\omega_i^{\star}-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)^2.\end{aligned}$$

To simplify, we further expand the leave-j-out propensity score, which satisfies

$$\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i=\hat{\mu}_i^{\star}-\hat{\mu}_i+\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^{\star}-r_j^{\star}),$$

hence

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)}-\hat{\boldsymbol{\theta}}\right) &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{a}}_i \\ &+ \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)+\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)^2 \\ &+ \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star) \\ &+ \frac{2\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_j^\star-r_j^\star) \\ &+ \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\hat{\mu}_j^\star-r_j^\star)^2.\end{aligned}$$

Note that

$$\begin{aligned}\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{a}}_i &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_i\hat{\mathbf{a}}_i\sum_j\left(\omega_j^\star(\omega_i^\star-\delta_{ij})\right) \\ &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_i\hat{\mathbf{a}}_i\left((n_\omega-\omega_i^\star)\omega_i^\star+\omega_i^\star(\omega_i^\star-1)\right) \\ &= \frac{\sqrt{n}}{n_\omega}\sum_i\omega_i^\star\hat{\mathbf{a}}_i.\end{aligned}$$

Similarly, we have

$$\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)=\frac{\sqrt{n}}{n_\omega}\sum_i\omega_i^\star\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right),$$

and

$$\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)^2=\frac{\sqrt{n}}{n_\omega}\sum_i\omega_i^\star\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)^2.$$

As a consequence,

$$\begin{aligned}(n_\omega-1)\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)}-\hat{\boldsymbol{\theta}}^\star\right) &= \frac{\sqrt{n}}{n_\omega}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star) \\ &+ \frac{2\sqrt{n}}{n_\omega}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_j^\star-r_j^\star) \\ &+ \frac{\sqrt{n}}{n_\omega}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\hat{\mu}_j^\star-r_j^\star)^2 \\ &= \frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star) \qquad \text{(I)} \\ &+ \frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_j^\star-r_j^\star) \qquad \text{(II)} \\ &+ \frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\hat{\mu}_j^\star-r_j^\star)^2 \qquad \text{(III)} \\ &+ o_{\mathbb{P}}(1).\end{aligned}$$

Next we analyze each term. For term (I), it is

$$\text{(I)}=\frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star)=\frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\sum_\ell\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell-e_j^\star\hat{\varepsilon}_j\Big)$$

$$= \frac{1}{\sqrt{n}} \sum_{i,j,\ell} \omega_j^\star (\omega_i^\star - \delta_{ij}) e_\ell^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{j\ell} \hat{\varepsilon}_\ell \tag{I.1}$$

$$- \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star e_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j. \tag{I.2}$$

Again we consider conditional expectation:

$$\begin{aligned}
\mathbb{E}^\star[(\mathrm{I}.1)] &= \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,\ell} \omega_j^\star (\omega_i^\star - \delta_{ij}) e_\ell^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{j\ell} \hat{\varepsilon}_\ell \right] \\
&= \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \right] + \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_j^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] \\
&+ \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i} \omega_i^\star (\omega_i^\star - 1) e_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ii}^2}{1-\pi_{ii}} \hat{\varepsilon}_i \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i + \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} \hat{\varepsilon}_j + \frac{1}{\sqrt{n}} \sum_i (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}^2}{1-\pi_{ii}} \hat{\varepsilon}_i.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}^\star[(\mathrm{I}.2)] &= \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star e_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] \\
&= \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star e_j^\star \omega_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] + \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i} \omega_i^\star e_i^\star (\omega_i^\star - 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \hat{\varepsilon}_i \right] \\
&= -\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j - \frac{1}{\sqrt{n}} \sum_i (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \hat{\varepsilon}_i.
\end{aligned}$$

Therefore

$$\mathbb{E}^\star[(\mathrm{I})] = \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \tag{I.3}$$

$$- \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \pi_{ij} \hat{\varepsilon}_j \tag{I.4}$$

$$- \frac{1}{\sqrt{n}} \sum_i (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \pi_{ii} \hat{\varepsilon}_i. \tag{I.5}$$

Furthermore,

$$\begin{aligned}
(\mathrm{I}.3) &= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \varepsilon_i + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \left(\pi_{ij}^2 + \frac{\pi_{ij}^2 \pi_{jj}}{1-\pi_{jj}} \right) \varepsilon_i + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \pi_{ij}^2 \varepsilon_i + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_i \pi_{ij}^2 \varepsilon_i + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_i \pi_{ii} \varepsilon_i + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\mathbf{b}_i \varepsilon_i | \mathbf{z}_i] \pi_{ii} + o_{\mathbb{P}}(1) \\
&= \mathbf{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \pi_{ii} + o_{\mathbb{P}}(1).
\end{aligned}$$

The second line follows from consistency and (E.35); the third line follows from Assumption A.3(2) and (E.36); the

fourth line is a simple fact of Lemma SA.3. Similar argument applies to (I.5), which implies

$$\text{(I.5)} = -\frac{1}{\sqrt{n}}\sum_i \mathbf{\Sigma}_0(\mathbb{E}^\star[e_i^{\star 3}]+1)\mathbf{b}_{1,i}\pi_{ii} + o_{\mathbb{P}}(1).$$

Finally,

$$\begin{aligned}\text{(I.4)} &= -\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{b}}_i\pi_{ij}\hat{\varepsilon}_j = -\frac{1}{\sqrt{n}}\sum_{i,j}\hat{\mathbf{b}}_i\pi_{ij}\hat{\varepsilon}_j + \frac{1}{\sqrt{n}}\sum_i\hat{\mathbf{b}}_i\pi_{ii}\hat{\varepsilon}_i \\ &= \frac{1}{\sqrt{n}}\sum_i\hat{\mathbf{b}}_i\pi_{ii}\hat{\varepsilon}_i = \frac{1}{\sqrt{n}}\sum_i\mathbf{\Sigma}_0\mathbf{b}_{1,i}\pi_{ii} + o_{\mathbb{P}}(1),\end{aligned}$$

where, in the second line, we used the fact that $\sum_{ij}\pi_{ij}\hat{\varepsilon}_j = 0$ for all i. Therefore

$$\text{(I)} = (1-\mathbb{E}^\star[e_i^{\star 3}])\frac{1}{\sqrt{n}}\sum_i\mathbf{\Sigma}_0\mathbf{b}_{1,i}\pi_{ii} + o_{\mathbb{P}}(1).$$

Next we consider (II). Note that it has the expansion:

$$\begin{aligned}\text{(II)} &= \frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_j^\star-r_j^\star) \\ &= \frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\sum_\ell\pi_{i\ell}e_\ell^\star\hat{\varepsilon}_\ell)(\sum_\ell\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell - e_j^\star\hat{\varepsilon}_j) \\ &= \frac{2}{\sqrt{n}}\sum_{i,j,\ell,\ell'}\omega_j^\star(\omega_i^\star-\delta_{ij})e_\ell^\star e_{\ell'}^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{i\ell}\pi_{j\ell'}\hat{\varepsilon}_\ell\hat{\varepsilon}_{\ell'} \qquad \text{(II.1)} \\ &\quad - \frac{2}{\sqrt{n}}\sum_{i,j,\ell}\omega_j^\star e_j^\star(\omega_i^\star-\delta_{ij})e_\ell^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{i\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_j. \qquad \text{(II.2)}\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}^\star[\text{(II.1)}] &= \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_i\omega_i^\star(\omega_i^\star-1)e_i^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ii}}{1-\pi_{ii}}\pi_{ii}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] \\ &+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_i^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ii}\pi_{ji}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] \\ &+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_j^\star e_j^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ij}\pi_{jj}\hat{\varepsilon}_j\hat{\varepsilon}_j\right] \\ &+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,\ell,i\neq\ell}\omega_i^\star(\omega_i^\star-1)e_\ell^\star e_\ell^\star\hat{\mathbf{c}}_i\frac{\pi_{ii}}{1-\pi_{ii}}\pi_{i\ell}\pi_{i\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_\ell\right] \\ &+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_i^\star e_j^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ii}\pi_{jj}\hat{\varepsilon}_i\hat{\varepsilon}_j\right] \\ &+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_j^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ij}\pi_{ji}\hat{\varepsilon}_j\hat{\varepsilon}_i\right] \\ &+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\omega_j^\star(\omega_i^\star-\delta_{ij})e_\ell^\star e_\ell^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{i\ell}\pi_{j\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_\ell\right] \\ &= \frac{1}{\sqrt{n}}O_{\mathbb{P}}\left(\sum_i\pi_{ii}^3 + \sum_{i,j}\pi_{ij}^2\pi_{ii} + \sum_{i,j}\pi_{ij}^2\pi_{jj} + \sum_{i,\ell}\pi_{i\ell}^2\pi_{ii}\right)\end{aligned}$$

$$+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2$$
$$=\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2+o_{\mathbb{P}}(1),$$

where the $o_{\mathbb{P}}(1)$ terms follows from (E.36) and Assumption A.3(1). Similarly,

$$\begin{aligned}
\mathbb{E}^\star[(\text{II}.2)] &= \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_i\omega_i^\star e_i^\star(\omega_i^\star-1)e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ii}}{1-\pi_{ii}}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_i\right]\\
&+\mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star e_j^\star\omega_i^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_j\right]\\
&+\mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star e_j^\star\omega_i^\star e_j^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ij}\hat{\varepsilon}_j\hat{\varepsilon}_j\right]\\
&=\frac{1}{\sqrt{n}}O_{\mathbb{P}}(\sum_i\pi_{ii}^2)-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2\\
&=-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2+o_{\mathbb{P}}(1).
\end{aligned}$$

Hence

$$\mathbb{E}^\star[(\text{II})]=\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j \tag{II.3}$$
$$+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j \tag{II.4}$$
$$+\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2 \tag{II.5}$$
$$-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j \tag{II.6}$$
$$-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2+o_{\mathbb{P}}(1). \tag{II.7}$$

First note that

$$(\text{II}.4)=\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbf{c}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\varepsilon_i\varepsilon_j+o_{\mathbb{P}}(1)=\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}[\mathbf{c}_i\varepsilon_i\varepsilon_j|\mathbf{z}_i,\mathbf{z}_j]\frac{\pi_{ij}^3}{1-\pi_{jj}}+o_{\mathbb{P}}(1)=o_{\mathbb{P}}(1).$$

Next

$$\begin{aligned}
(\text{II}.3)+(\text{II}.6)&=-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\pi_{ij}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_j=-\frac{2}{\sqrt{n}}\sum_{i,j}\hat{\mathbf{c}}_i\pi_{ij}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_i\hat{\mathbf{c}}_i\pi_{ii}^2\hat{\varepsilon}_i^2\\
&=\frac{2}{\sqrt{n}}\sum_i\hat{\mathbf{c}}_i\pi_{ii}^2\hat{\varepsilon}_i^2=o_{\mathbb{P}}(1),
\end{aligned}$$

where for the third line we used the fact $\sum_{i,j}\pi_{ij}\hat{\varepsilon}_j=0$, and the last line follows from Assumption A.3(1). Hence

$$\mathbb{E}^\star[(\text{II})]=\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2 \tag{II.8}$$
$$-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}^\star[e_i^{\star 3}]\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2+o_{\mathbb{P}}(1). \tag{II.9}$$

For the first line, we have the following result:

$$
\begin{aligned}
\text{(II.8)} &= \left| \frac{2}{\sqrt{n}} \sum_{\substack{i,j,\ell \\ \text{distinct}}} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{c}}_i \hat{\varepsilon}_j^2 \frac{\pi_{ij}^2}{1-\pi_{jj}} \right| \\
&= \left| \frac{2}{\sqrt{n}} \sum_{\substack{i,j,\ell \\ \text{distinct}}} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,\ell,i\neq \ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{i\ell}^2}{1-\pi_{\ell\ell}} \right| && \text{(change } j \to \ell\text{)} \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,\ell,i\neq \ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{i\ell}^2}{1-\pi_{\ell\ell}} \right| + o_{\mathbb{P}}(1) && \text{((E.36) and Assumption A.3(1))} \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,\ell,i\neq \ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \pi_{i\ell}^2 \right| + o_{\mathbb{P}}(1) && \text{((E.36) and Assumption A.3(2))} \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,\ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \pi_{i\ell}^2 \right| + o_{\mathbb{P}}(1) && \text{(Assumption A.3(1))} \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \pi_{ij}\pi_{i\ell}\pi_{j\ell} \right| + o_{\mathbb{P}}(1) = \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \hat{\mathbf{c}}_i \hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}} \right| + o_{\mathbb{P}}(1) \\
&\lesssim_{\mathbb{P}} \frac{1}{\sqrt{n}} \sqrt{\sum_{i,\ell} \pi_{i\ell}^2} \sqrt{\sum_{i,\ell} \left(\sum_j \frac{\pi_{ij}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}} \right)^2} = \frac{\sqrt{k}}{\sqrt{n}} \sqrt{\sum_{i,\ell} \left(\sum_j \frac{\pi_{ij}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}} \right)^2} \\
&= \frac{\sqrt{k}}{\sqrt{n}} \sqrt{\sum_{i,\ell} \sum_{jj'} \frac{\pi_{ij}\pi_{j\ell}\pi_{jj}\pi_{ij'}\pi_{j'\ell}\pi_{j'j'}}{(1-\pi_{jj})(1-\pi_{j'j'})}} = \frac{\sqrt{k}}{\sqrt{n}} \sqrt{\sum_{jj'} \frac{\pi_{jj}\pi_{j'j'}\pi_{jj'}^2}{(1-\pi_{jj})(1-\pi_{j'j'})}} \\
&\lesssim_{\mathbb{P}} \frac{\sqrt{k}}{\sqrt{n}} \sqrt{\sum_{jj'} \pi_{jj}\pi_{j'j'}\pi_{jj'}^2} = \frac{\sqrt{k}}{\sqrt{n}} \cdot o_{\mathbb{P}}(\sqrt{k}) = o_{\mathbb{P}}(1).
\end{aligned}
$$

Hence we have:

$$
\begin{aligned}
\text{(II)} &= -\frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbb{E}^\star[e_i^{\star 3}] \hat{\mathbf{c}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_j^2 + o_{\mathbb{P}}(1) = -\frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbb{E}^\star[e_i^{\star 3}] \mathbf{c}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \varepsilon_j^2 + o_{\mathbb{P}}(1) \\
&= -\frac{2}{\sqrt{n}} \sum_{i,j} \mathbb{E}^\star[e_i^{\star 3}] \mathbf{c}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \varepsilon_j^2 + o_{\mathbb{P}}(1) = -\mathbb{E}^\star[e_i^{\star 3}] \mathbf{\Sigma}_0 \frac{2}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2 + o_{\mathbb{P}}(1),
\end{aligned}
$$

and the last line follows essentially from Lemma SA.4.

(III) has the following expansion:

$$
\begin{aligned}
\text{(III)} &= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\hat{\mu}_j^\star - r_j^\star)^2 \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell - e_j^\star \hat{\varepsilon}_j)^2 \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell)^2 && \text{(III.1)} \\
&\quad - \frac{2}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell) e_j^\star \hat{\varepsilon}_j && \text{(III.2)} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (e_j^\star \hat{\varepsilon}_j)^2. && \text{(III.3)}
\end{aligned}
$$

Then

$$\begin{aligned}
\mathbb{E}^\star[(\text{III.1})] &= \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\sum_\ell \pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell\right)^2\right] \\
&= \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i}\omega_i^\star(\omega_i^\star-1)e_i^\star e_i^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ii}}{1-\pi_{ii}}\right)^2\pi_{ii}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] \\
&+ \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_i^\star e_i^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ji}\pi_{ji}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] \\
&+ \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_j^\star e_j^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{jj}\pi_{jj}\hat{\varepsilon}_j\hat{\varepsilon}_j\right] \\
&+ \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i,\ell,i\neq \ell}\omega_i^\star(\omega_i^\star-1)e_\ell^\star e_\ell^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ii}}{1-\pi_{ii}}\right)^2\pi_{i\ell}\pi_{i\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_\ell\right] \\
&+ \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star(\omega_i^\star-\delta_{ij})e_i^\star e_j^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ji}\pi_{jj}\hat{\varepsilon}_i\hat{\varepsilon}_j\right] \\
&+ \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star(\omega_i^\star-\delta_{ij})e_j^\star e_i^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{jj}\pi_{ji}\hat{\varepsilon}_j\hat{\varepsilon}_i\right] \\
&+ \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\omega_j^\star(\omega_i^\star-\delta_{ij})e_\ell^\star e_\ell^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{j\ell}\pi_{j\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_\ell\right] \\
&= \frac{1}{\sqrt{n}}O_{\mathbb{P}}\left(\sum_i\pi_{ii}^4+\sum_{i,j}\pi_{ij}^4+\sum_{i,j}\pi_{ij}^2\pi_{jj}^2+\sum_{i,\ell}\pi_{i\ell}^2\pi_{ii}^2+\sum_{i,j}\pi_{ij}^3\pi_{jj}+\sum_{i,j,\ell}\pi_{ij}^2\pi_{j\ell}^2\right) \\
&= o_{\mathbb{P}}(1),
\end{aligned}$$

by (E.36), (E.37) and Assumption A.3(1). Next

$$\begin{aligned}
\mathbb{E}^\star[(\text{III.2})] &= \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\sum_\ell\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell\right)e_j^\star\hat{\varepsilon}_j\right] \\
&= \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i}\omega_i^\star e_i^\star(\omega_i^\star-1)e_i^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ii}}{1-\pi_{ii}}\right)^2\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] \\
&+ \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star e_j^\star\omega_i^\star e_i^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ji}\hat{\varepsilon}_i\hat{\varepsilon}_j\right] \\
&+ \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star e_j^\star\omega_i^\star e_j^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{jj}\hat{\varepsilon}_j\hat{\varepsilon}_j\right] \\
&= -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ij}\hat{\varepsilon}_i\hat{\varepsilon}_j+o_{\mathbb{P}}(1) = -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ij}\varepsilon_i\varepsilon_j+o_{\mathbb{P}}(1) \\
&= -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}[\mathbf{c}_i\varepsilon_i\varepsilon_j|\mathbf{z}_i,\mathbf{z}_j]\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ij}+o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).
\end{aligned}$$

Finally

$$\mathbb{E}^\star[(\text{III.3})] = \frac{1}{\sqrt{n}}\sum_{i,j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\hat{\varepsilon}_j^2+o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}}\sum_{i,j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\varepsilon_j^2+o_{\mathbb{P}}(1)$$

$$= (\mathbb{E}^\star[e_i^{\star 3}] + 1)\boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij}\pi_{ij}^2 + o_{\mathbb{P}}(1).$$

Given the previous results,

$$(n_\omega - 1)\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}^\star\right) = (1 - \mathbb{E}^\star[e_i^{\star 3}])\boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left(\sum_i \mathbf{b}_{1,i}\pi_{ii} + \sum_{i,j}\mathbf{b}_{2,ij}\pi_{ij}^2\right) + o_{\mathbb{P}}(1)$$
$$= (1 - \mathbb{E}^\star[e_i^{\star 3}])\boldsymbol{\mathcal{B}} + o_{\mathbb{P}}(1).$$

■

SA-9.19.2 Part 2

We follow the notational convention used in the previous part:

$$\hat{\mathbf{a}}_i = \boldsymbol{\Sigma}_0 \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{b}}_i = \boldsymbol{\Sigma}_0 \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{c}}_i = \boldsymbol{\Sigma}_0 \frac{\ddot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})}{2}.$$

Similarly,

$$\mathbf{a}_i = \boldsymbol{\Sigma}_0 \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \qquad \mathbf{b}_i = \boldsymbol{\Sigma}_0 \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \qquad \mathbf{c}_i = \boldsymbol{\Sigma}_0 \frac{\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)}{2}.$$

First note that the jackknife variance estimator for the bootstrap data takes the form:

$$(n-1)\sum_j \left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)}\right)^2,$$

where for a (column) vector $\mathbf{v}$, we use $\mathbf{v}^2$ to denote $\mathbf{v}\mathbf{v}^\mathsf{T}$ to save space. Then the variance estimator could be rewritten as

$$\hat{\boldsymbol{\mathcal{V}}}^\star = (n-1)\sum_j \left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^\star\right)^2 - \frac{1}{n-1}\left(\hat{\boldsymbol{\mathcal{B}}}^\star\right)^2$$
$$= (n-1)\sum_j \left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^\star\right)^2 + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Next recall that

$$\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}} = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{a}}_i + \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right)$$
$$+ \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right)^2.$$

Then we make the following decomposition:

$$\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{a}}_i = \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{a}}_i - \frac{1}{n_\omega - 1}\hat{\mathbf{a}}_j,$$

and

$$\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right) = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i - \frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)$$
$$= \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)$$
$$= \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j\left(\hat{\mu}_j^\star - \hat{\mu}_j\right) - \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right),$$

and

$$\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right)^2 = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i - \frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)^2$$

$$
\begin{aligned}
&= \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&+ \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1} \sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&- \frac{1}{n_\omega - 1} \hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&+ \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right).
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}} &= \frac{1}{n_\omega - 1} \sum_i \omega_i^\star \hat{\mathbf{a}}_i \\
&- \frac{1}{n_\omega - 1} \hat{\mathbf{a}}_j \\
&+ \frac{1}{n_\omega - 1} \sum_i \omega_i^\star \hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \\
&- \frac{1}{n_\omega - 1} \hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) \\
&- \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&+ \frac{1}{n_\omega - 1} \sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&- \frac{1}{n_\omega - 1} \hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&+ \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}} &= \frac{1}{n_\omega} \sum_j \omega_j^\star \left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}\right) \\
&= \frac{1}{n_\omega - 1} \sum_i \omega_i^\star \hat{\mathbf{a}}_i \\
&- \frac{1}{n_\omega (n_\omega - 1)} \sum_j \omega_j^\star \hat{\mathbf{a}}_j \\
&+ \frac{1}{n_\omega - 1} \sum_i \omega_i^\star \hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \\
&- \frac{1}{n_\omega (n_\omega - 1)} \sum_j \omega_i^\star \hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) \\
&- \frac{1}{n_\omega (n_\omega - 1)} \sum_{i,j} (\omega_i^\star - \delta_{ij}) \omega_j^\star \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{n_\omega - 1} \sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&- \frac{1}{n_\omega(n_\omega - 1)} \sum_j \omega_j^\star \hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&+ \frac{1}{n_\omega(n_\omega - 1)} \sum_{i,j} (\omega_i^\star - \delta_{ij}) \omega_j^\star \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega(n_\omega - 1)} \sum_{i,j} (\omega_i^\star - \delta_{ij}) \omega_j^\star \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right),
\end{aligned}
$$

which means

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)} &= \frac{1}{n_\omega - 1} \left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mathcal{B}}}^\star / \sqrt{n_\omega}\right) - \frac{1}{n_\omega - 1} \hat{\mathbf{a}}_j - \frac{1}{n_\omega - 1} \hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) - \frac{1}{n_\omega - 1} \hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&- \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) + \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1} \left(\hat{\boldsymbol{\theta}}_{\mathsf{bc}}^\star - \hat{\boldsymbol{\theta}}\right) && \text{(I)} \\
&- \frac{1}{n_\omega - 1} \hat{\mathbf{a}}_j && \text{(II)} \\
&- \frac{1}{n_\omega - 1} \hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) && \text{(III)} \\
&- \frac{1}{n_\omega - 1} \hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 && \text{(IV)} \\
&- \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) && \text{(V)} \\
&+ \frac{1}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 && \text{(VI)} \\
&- \frac{2}{n_\omega - 1} \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right). && \text{(VII)}
\end{aligned}
$$

Term (I) is the easiest:

$$
(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{I})^2 \asymp \left(\hat{\boldsymbol{\theta}}_{\mathsf{bc}}^\star - \hat{\boldsymbol{\theta}}\right)^2 = o_{\mathbb{P}}(1),
$$

by consistency. Similarly

$$
(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{I}) \Big((\mathrm{II}) + \cdots (\mathrm{VII})\Big)^\mathsf{T} = \left(\hat{\boldsymbol{\theta}}_{\mathsf{bc}}^\star - \hat{\boldsymbol{\theta}}\right) \sum_j \omega_j^\star \Big((\mathrm{II}) + \cdots (\mathrm{VII})\Big)^\mathsf{T} = o_{\mathbb{P}}(1).
$$

Next

$$
(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{II})^2 = \frac{1}{n_\omega - 1} \sum_j \omega_j^\star \left(\hat{\mathbf{a}}_j\right)^2 \to_{\mathbb{P}} \mathbb{V}[\bar{\boldsymbol{\Psi}}_1].
$$

By the uniform consistency of $\hat{\mu}_j^\star$, it is very easy to show that

$$
(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{II})(\mathrm{III})^\mathsf{T} = o_{\mathbb{P}}(1), \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{II})(\mathrm{IV})^\mathsf{T} = o_{\mathbb{P}}(1).
$$

Then

$$
(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{II})(\mathrm{V})^\mathsf{T} = \frac{1}{n_\omega - 1} \sum_{i,j} \hat{\mathbf{a}}_j \hat{\mathbf{b}}_i^\mathsf{T} \omega_j^\star (\omega_i^\star - \delta_{ij}) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)
$$

$$
\begin{aligned}
&= \frac{1}{n_\omega - 1} \sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_i \left[\hat{\mathbf{b}}_i^\mathsf{T} (\omega_i^\star - \delta_{ij}) \frac{\pi_{ij}}{1 - \pi_{jj}} \right] \\
&= \frac{1}{n_\omega - 1} \sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_i \left[\hat{\mathbf{b}}_i^\mathsf{T} \pi_{ij} \right] && \text{(i)} \\
&+ \frac{1}{n_\omega - 1} \sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_i \left[\hat{\mathbf{b}}_i^\mathsf{T} \frac{\pi_{ij} \pi_{jj}}{1 - \pi_{jj}} \right] && \text{(ii)} \\
&+ \frac{1}{n_\omega - 1} \sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_{i, i \neq j} \left[\hat{\mathbf{b}}_i^\mathsf{T} e_i^\star \frac{\pi_{ij}}{1 - \pi_{jj}} \right] && \text{(iii)} \\
&+ \frac{1}{n_\omega - 1} \sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \left[\hat{\mathbf{b}}_j^\mathsf{T} (e_j^\star - 1) \frac{\pi_{jj}}{1 - \pi_{jj}} \right]. && \text{(iv)}
\end{aligned}
$$

Then we have (i) $\to_\mathbb{P} \mathbb{C}\mathrm{ov}[\bar{\mathbf{\Psi}}_1, \bar{\mathbf{\Psi}}_2 | \mathbf{Z}]$, and the other terms are asymptotically negligible. This essentially uses the same technique (conditional mean and variance calculation) used for Lemma SA.3 and SA.4, and we do not repeat here. By taking transpose, we have $(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{V})(\mathrm{II})^\mathsf{T} \to_\mathbb{P} \mathbb{C}\mathrm{ov}[\bar{\mathbf{\Psi}}_2, \bar{\mathbf{\Psi}}_1 | \mathbf{Z}]$. Further,

$$
\begin{aligned}
\left| (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{II})(\mathrm{VI})^\mathsf{T} \right| &= \left| \frac{1}{n_\omega - 1} \sum_j \omega_j^\star \hat{\mathbf{a}}_j \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}} \right)^2 \left(e_j^\star \hat{\varepsilon}_j \right)^2 \right| \\
&\precsim_\mathbb{P} \frac{1}{n} \sum_{i,j} \pi_{ij}^2 = o_\mathbb{P}(1),
\end{aligned}
$$

and

$$
\begin{aligned}
\left| (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{II})(\mathrm{VII})^\mathsf{T} \right| &= \left| \frac{2}{n_\omega - 1} \sum_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \hat{\mathbf{a}}_j \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i \right) \left(\frac{\pi_{ij}}{1 - \pi_{jj}} \right) \right| \\
&\precsim_\mathbb{P} \frac{1}{n} \cdot \sqrt{\sum_j |\omega_j^\star e_j^\star \hat{\varepsilon}_j \hat{\mathbf{a}}_j|^2} \sqrt{\sum_j \left| (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i \right) \right|^2} \\
&= o_\mathbb{P}(1).
\end{aligned}
$$

Due to uniform consistency of $\hat{\mu}_j^\star$, the following are easy to establish:

$$
\begin{aligned}
&(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{III})^2 = o_\mathbb{P}(1) \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{III})(\mathrm{IV})^\mathsf{T} = o_\mathbb{P}(1) \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{III})(\mathrm{V})^\mathsf{T} = o_\mathbb{P}(1) \\
&(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{III})(\mathrm{VI})^\mathsf{T} = o_\mathbb{P}(1) \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{III})(\mathrm{VII})^\mathsf{T} = o_\mathbb{P}(1),
\end{aligned}
$$

as well as

$$
\begin{aligned}
&(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{IV})^2 = o_\mathbb{P}(1) \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{IV})(\mathrm{V})^\mathsf{T} = o_\mathbb{P}(1) \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{IV})(\mathrm{VI})^\mathsf{T} = o_\mathbb{P}(1) \\
&(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{IV})(\mathrm{VII})^\mathsf{T} = o_\mathbb{P}(1).
\end{aligned}
$$

Next it is easy to show that

$$
(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{V})^2 \to_\mathbb{P} (1 + \mathbb{E}^\star[e_i^{\star 3}]) \mathbb{V}[\bar{\mathbf{\Psi}}_2 | \mathbf{Z}].
$$

What remains are terms involving $(\mathrm{V})(\mathrm{VI})^\mathsf{T}$, $(\mathrm{V})(\mathrm{VII})^\mathsf{T}$, $(\mathrm{VI})^2$, $(\mathrm{VI})(\mathrm{VII})^\mathsf{T}$ and $(\mathrm{VII})^2$.

$$
\begin{aligned}
&\left| (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{V})(\mathrm{VI})^\mathsf{T} \right| \\
&= \left| \frac{1}{n_\omega - 1} \sum_j \omega_j^\star \left(\sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j \right) \right) \left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j}) \hat{\mathbf{c}}_\ell \left(\frac{\pi_{\ell j}}{1 - \pi_{jj}} \right)^2 \left(e_j^\star \hat{\varepsilon}_j \right)^2 \right)^\mathsf{T} \right|
\end{aligned}
$$

$$
\begin{aligned}
&\asymp_{\mathbb{P}} \left| \frac{1}{n_\omega - 1} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \left(e_j^\star \hat{\varepsilon}_j \right)^3 \pi_{ij} \left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j}) \hat{\mathbf{c}}_\ell \left(\frac{\pi_{\ell j}}{1 - \pi_{jj}} \right)^2 \right)^\mathsf{T} \right| \\
&\precsim_{\mathbb{P}} \left(\frac{1}{n} \sum_j \left| \sum_\ell (\omega_\ell^\star - \delta_{\ell j}) \hat{\mathbf{c}}_\ell \left(\frac{\pi_{\ell j}}{1 - \pi_{jj}} \right)^2 \right|^2 \right)^{1/2} \\
&\asymp_p \sqrt{\frac{1}{n} \sum_{j,i,\ell} \pi_{ij}^2 \pi_{\ell j}^2} = o_{\mathbb{P}}(1).
\end{aligned}
$$

And

$$
\begin{aligned}
&\left| (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{V}) (\mathrm{VII})^\mathsf{T} \right| \\
&= \left| \frac{2}{n_\omega - 1} \sum_j \left(\sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j \right) \right) \left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j}) \hat{\mathbf{c}}_\ell \left(\hat{\mu}_\ell^\star - \hat{\mu}_\ell \right) \left(\frac{\pi_{\ell j}}{1 - \pi_{jj}} e_j^\star \hat{\varepsilon}_j \right) \right)^\mathsf{T} \right| \\
&= \left| \frac{2}{n_\omega - 1} \sum_{i,j} (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1 - \pi_{jj}} (e_j^\star \hat{\varepsilon}_j)^2 \left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j}) \hat{\mathbf{c}}_\ell \left(\hat{\mu}_\ell^\star - \hat{\mu}_\ell \right) \left(\frac{\pi_{\ell j}}{1 - \pi_{jj}} \right) \right)^\mathsf{T} \right| \\
&\precsim_{\mathbb{P}} \sqrt{\frac{1}{n} \sum_j \left| \sum_\ell \frac{\pi_{\ell j}}{1 - \pi_{\ell\ell}} \left(\hat{\mu}_\ell^\star - \hat{\mu}_\ell \right) \right|^2} = o_{\mathbb{P}}(1),
\end{aligned}
$$

Using techniques in the above results, we can show

$$
(n_\omega - 1) \sum_j \omega_j^\star (\mathrm{VI})^2 = o_{\mathbb{P}}(1), \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{VII})^2 = o_{\mathbb{P}}(1), \qquad (n_\omega - 1) \sum_j \omega_j^\star (\mathrm{VI})(\mathrm{VII})^\mathsf{T} = o_{\mathbb{P}}(1),
$$

which closes the proof. ■

SA-10 Empirical Papers with Possibly Many Covariates

Per request of the Editor and the Reviewers, we document a sample of empirical papers employing two-step estimation strategies where the dimensionality of the covariates used is possibly "large" (in the sense that $k/\sqrt{n}$ is large) and therefore our methods could have been used to obtain more robust inference procedures.

These papers were found upon searching for the following keywords: "propensity score", "control function", and "semiparametric". We only report those papers that explicitly declare the dimensionality of the first step estimation and exclude those papers that did not provide this information clearly (even though these also appear to be using several covariates and/or transformations thereof). This list is not meant to be systematic or exhaustive, and therefore we did not attempt to conduct a meta-analysis on the topic of many covariates in two-step estimation.

1. Abadie (2003), *Journal of Econometrics* 113(2): 231–263.

   Methodology: local average response function method. 86 covariates are used to estimate a linear probability model with $n = 9,275$; $k/\sqrt{n} \geq 0.89$, depending of specification considered.

2. Black and Smith (2004), *Journal of Econometrics* 121(1): 99–124.

   Methodology: propensity score matching. More than 30 covariates are used in propensity score estimation with $n \approx 350$; $k/\sqrt{n} \geq 1.60$, depending of specification considered.

3. Brand and Davis (2011), *Demography* 48(3): 863–887.

   Methodology: propensity score is estimated with Probit model, which is then used as generated regressor. 20 covariates are used for propensity score estimation with $n \approx 2,000$; $k/\sqrt{n} \geq 0.60$, depending of specification considered.

4. Brand and Xie (2010), *American Sociological Review* 75(2): 273–302.

   Methodology: propensity score is estimated with Logit model, which is then used to formed strata for treatment effect estimation. About 18 covariates are used for propensity score estimation with $n \approx 1,250$; $k/\sqrt{n} \geq 0.50$, depending of specification considered.

5. Carneiro, Heckman and Vytlacil (2011), *American Economic Review* 101(6): 2754–2781.

   Methodology: propensity score is estimated with Logit model and the fitted value is used as generated regressor to estimate a partially linear second step. 35 covariates are used to estimate the propensity score with sample size $n = 1,747$; $k/\sqrt{n} \geq 0.85$, depending of specification considered.

6. Galasso and Schankerman (2014), *Quarterly Journal of Economics* 130(1): 317–369.

   Methodology: predicted probability is used as instrument. 51 fixed effects plus other variables are used in estimating the conditional probability, with sample size $n = 1,357$; $k/\sqrt{n} \geq 1.38$, depending of specification considered.

7. Helpman, Melitz and Rubinstein (2008), *Quarterly Journal of Economics*, 123(2): 441–487.

   Methodology: probability of exports is estimated in the first step and then used as generated regressor. About 340 covariates are used with sample size $n \approx 24,700$; $k/\sqrt{n} \geq 2.04$, depending of specification considered.

8. Jalan and Ravallion (2003), *Journal of Econometrics* 112(1): 153–173.

   Methodology: propensity score matching method. 91 covariates are used in propensity score matching with $n \approx 30,000$; $k/\sqrt{n} \approx 0.60$, depending of specification considered.

9. Lechner (1999), *Journal of Business & Economic Statistics* 17(1): 74–90.

   Methodology: propensity score matching. 31 covariates are used with sample size $n = 1,399$; $k/\sqrt{n} \geq 0.85$, depending of specification considered.

10. Lechner and Wunsch (2013), *Journal of Econometrics* 21: 111–121.

    Methodology: propensity score is estimated with Probit model, and is used for treatment effect estimation (simulation). More than 180 covariates are used with $n \approx 25,000$; $k/\sqrt{n} \geq 1.10$, depending of specification considered.

11. Noboa-Hidalgo and Urzúa (2012), *Journal of Human Capital* 6(1): 1–34.

    Methodology: propensity score is estimated with Probit model, which is then used as generated regressor. About 20 covariates are used for propensity score estimation with $n = 469$; $k/\sqrt{n} \geq 0.90$, depending of specification considered.

12. Olley and Pakes (1996), *Econometrica* 64(6): 1263–1297.

    Methodology: three-step procedure described in Section SA-5.7. Fourth-order series expansion of three variables are used in the first step (34 covariates), and $n \approx 1,000$; $k/\sqrt{n} \geq 1.00$, depending of specification considered.

13. Tsai and Xie (2011), *Social Science Research* 40(3): 796–810.

    Methodology: propensity score is estimated with Probit model. 34 covariates are used with sample size $n \approx 1,300$; $k/\sqrt{n} \geq 0.85$, depending of specification considered.

In this sample of papers, we found that $k/\sqrt{n}$ is roughly around 1.00. According to our simulations, which employed a very simple data generating process, two-step conventional inference procedures constructed using $k/\sqrt{n} \approx 1.00$ exhibit an empirical size distortion of about $10-15$ percentage points. That is, a nominal 95% conventional confidence interval exhibits empirical coverage of about $80-85\%$.

References

Abadie, Alberto. (2003). 'Semiparametric Instrumental Variable Estimation of Treatment Response Models', *Journal of Econometrics* 113(2), 231–263.

Abadie, Alberto. (2005). 'Semiparametric Difference-in-Differences Estimators', *Review of Economic Studies* 72(1), 1–19.

Abadie, Alberto and Cattaneo, Matias D. (2018). 'Econometric Methods for Program Evaluation', *Annual Review of Economics* 10, 465–503.

Belloni, Alexandre, Chernozhukov, Victor, Chetverikov, Denis and Kato, Kengo. (2015). 'Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results', *Journal of Econometrics* 186(2), 345–366.

Carneiro, Pedro, Heckman, James J. and Vytlacil, Edward J. (2011). 'Estimating Marginal Returns to Education', *American Economic Review* 101(6), 2754–2781.

Cattaneo, Matias D., Farrell, Max H. and Feng, Yingjie. (2018). 'Large Sample Properties of Partitioning-Based Estimators', arXiv:1804.04916 .

Cattaneo, Matias D., Jansson, Michael and Newey, Whitney K. (2018*a*). 'Alternative Asymptotics and the Partially Linear Model with Many Regressors', *Econometric Theory* 34(2), 277–301.

Cattaneo, Matias D., Jansson, Michael and Newey, Whitney K. (2018*b*). 'Inference in Linear Regression Models with Many Covariates and Heteroskedasticity', *Journal of the American Statistical Association*, forthcoming .

Heckman, James J. and Vytlacil, Edward J. (2005). 'Structural Equations, Treatment Effects and Econometric Policy Evaluation', *Econometrica* 73(3), 669–738.

Imbens, Guido W., Angrist, Joshua D. and Krueger, Alan B. (1999). 'Jackknife Instrumental Variables Estimation', *Journal of Applied Econometrics* 14(1).

Olley, G. Steven and Pakes, Ariel. (1996). 'The Dynamics of Productivity in the Telecommunications Equipment Industry', *Econometrica* 64(6), 1263–1297.

van der Vaart, Aad W. and Wellner, Jon A. (1996), *Weak Convergence and Empirical Processes*, Springer, New York.

Vershynin, Roman. (2018), *High-Dimensional Probability*, Cambridge University Press, New York.

Wooldridge, Jeffrey M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2 edn, MIT Press, Cambridge, MA.

Wooldridge, Jeffrey M. (2015). 'Control Function Methods in Applied Econometrics', *Journal of Human Resources* 50(2), 420–445.

Table 1. Bootstrap Inference, MTE, DGP 1
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: conventional							$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: percentile ci						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.16	0.43	4.81	4.83	0.05	18.85	0.02	19.71	0.09	5.03	5.03	0.05	19.73	0.03	19.71
20	0.02	0.63	2.06	4.24	4.71	0.07	16.60	0.08	16.28	0.86	5.16	5.23	0.05	20.23	0.10	16.28
40	0.04	1.26	3.30	3.67	4.93	0.15	14.38	0.16	13.85	1.85	4.79	5.13	0.06	18.76	0.17	13.85
60	0.06	1.90	4.14	3.27	5.27	0.23	12.81	0.26	12.34	2.61	4.40	5.11	0.09	17.23	0.22	12.34
80	0.08	2.53	4.76	3.01	5.63	0.36	11.81	0.39	11.29	3.14	4.10	5.17	0.11	16.09	0.28	11.29
100	0.10	3.16	5.27	2.80	5.97	0.47	10.97	0.50	10.57	3.55	3.84	5.23	0.15	15.04	0.33	10.57
120	0.12	3.79	5.73	2.65	6.31	0.58	10.39	0.59	9.94	3.90	3.66	5.34	0.18	14.34	0.39	9.94
140	0.14	4.43	6.11	2.54	6.62	0.67	9.94	0.70	9.51	4.15	3.51	5.43	0.23	13.75	0.44	9.51
160	0.16	5.06	6.46	2.44	6.90	0.75	9.58	0.79	9.10	4.37	3.39	5.53	0.26	13.27	0.48	9.10
180	0.18	5.69	6.80	2.33	7.19	0.84	9.12	0.85	8.78	4.61	3.22	5.62	0.30	12.62	0.53	8.78
200	0.20	6.32	7.11	2.24	7.46	0.89	8.76	0.90	8.49	4.82	3.09	5.73	0.34	12.11	0.58	8.49

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: conventional							$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: percentile ci						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.11	0.46	4.78	4.80	0.05	18.73	0.04	18.94	0.21	4.88	4.89	0.05	19.14	0.05	18.94
20	0.01	0.45	1.69	4.43	4.75	0.07	17.37	0.07	17.32	0.51	5.03	5.05	0.05	19.70	0.09	17.32
40	0.02	0.89	3.03	4.03	5.05	0.12	15.80	0.13	15.64	1.35	4.90	5.08	0.06	19.22	0.12	15.64
60	0.03	1.34	3.97	3.81	5.50	0.18	14.95	0.20	14.37	2.07	4.81	5.24	0.07	18.86	0.18	14.37
80	0.04	1.79	4.75	3.58	5.95	0.27	14.04	0.30	13.44	2.76	4.63	5.39	0.09	18.13	0.22	13.44
100	0.05	2.24	5.37	3.37	6.34	0.35	13.21	0.39	12.70	3.32	4.42	5.53	0.11	17.34	0.26	12.70
120	0.06	2.68	5.88	3.21	6.70	0.46	12.59	0.49	12.08	3.76	4.27	5.69	0.14	16.74	0.32	12.08
140	0.07	3.13	6.35	3.14	7.08	0.54	12.32	0.57	11.57	4.18	4.21	5.93	0.17	16.51	0.37	11.57
160	0.08	3.58	6.77	3.02	7.41	0.62	11.83	0.65	11.15	4.53	4.08	6.10	0.21	16.00	0.42	11.15
180	0.09	4.02	7.15	2.94	7.73	0.68	11.51	0.71	10.73	4.84	3.99	6.28	0.23	15.65	0.46	10.73
200	0.10	4.47	7.47	2.83	7.99	0.75	11.10	0.78	10.40	5.07	3.86	6.38	0.26	15.15	0.50	10.40

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. $k = 5$ is the correctly specified model.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the bootstrap (500 repetitions, Rademacher weights). For the **naive ci**, we first center the bootstrap distribution to suppress its bias correction ability; **(viii)**: ci‡: average confidence interval length.

Table 2. Jackknife Inference, MTE, DGP 1
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.16	0.16	4.78	4.79	0.05	18.75	0.04	19.28	−0.20	5.00	5.00	0.05	19.60	0.05	19.28
20	0.02	0.63	1.79	4.16	4.53	0.07	16.32	0.05	18.29	0.22	5.34	5.34	0.06	20.93	0.08	18.29
40	0.04	1.26	3.08	3.70	4.82	0.12	14.52	0.07	17.06	0.96	5.42	5.51	0.06	21.25	0.11	17.06
60	0.06	1.90	3.95	3.30	5.15	0.23	12.92	0.12	15.93	1.68	5.19	5.46	0.06	20.35	0.14	15.93
80	0.08	2.53	4.64	3.04	5.55	0.34	11.91	0.18	15.00	2.30	5.04	5.54	0.08	19.75	0.18	15.00
100	0.10	3.16	5.14	2.81	5.86	0.45	11.02	0.24	14.24	2.69	4.84	5.53	0.08	18.96	0.20	14.24
120	0.12	3.79	5.65	2.63	6.23	0.58	10.29	0.33	13.61	3.13	4.70	5.64	0.10	18.40	0.23	13.61
140	0.14	4.43	6.05	2.51	6.55	0.67	9.86	0.43	13.10	3.42	4.55	5.69	0.11	17.84	0.24	13.10
160	0.16	5.06	6.39	2.42	6.83	0.76	9.47	0.51	12.66	3.50	4.49	5.70	0.12	17.62	0.27	12.66
180	0.18	5.69	6.77	2.32	7.15	0.83	9.08	0.60	12.24	3.72	4.41	5.77	0.14	17.30	0.31	12.24
200	0.20	6.32	7.13	2.24	7.47	0.89	8.76	0.68	11.92	3.94	4.34	5.86	0.15	17.00	0.33	11.92

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.11	0.33	4.73	4.74	0.05	18.54	0.04	18.84	0.08	4.83	4.83	0.06	18.94	0.04	18.84
20	0.01	0.45	1.65	4.37	4.68	0.06	17.15	0.05	18.48	0.30	5.07	5.08	0.05	19.89	0.06	18.48
40	0.02	0.89	2.94	4.08	5.03	0.10	15.99	0.07	17.96	0.77	5.27	5.32	0.05	20.64	0.08	17.96
60	0.03	1.34	3.93	3.84	5.49	0.17	15.05	0.11	17.34	1.35	5.33	5.49	0.05	20.88	0.10	17.34
80	0.04	1.79	4.76	3.61	5.98	0.26	14.16	0.16	16.74	1.98	5.24	5.61	0.07	20.56	0.13	16.74
100	0.05	2.24	5.42	3.40	6.40	0.36	13.33	0.22	16.22	2.54	5.13	5.72	0.08	20.10	0.16	16.22
120	0.06	2.68	5.95	3.24	6.78	0.45	12.71	0.29	15.69	2.98	5.05	5.86	0.09	19.78	0.19	15.69
140	0.07	3.13	6.38	3.08	7.08	0.55	12.08	0.35	15.27	3.32	4.93	5.94	0.10	19.33	0.19	15.27
160	0.08	3.58	6.76	2.98	7.39	0.62	11.70	0.43	14.83	3.60	4.85	6.04	0.12	19.00	0.23	14.83
180	0.09	4.02	7.14	2.91	7.71	0.69	11.42	0.49	14.45	3.95	4.84	6.25	0.13	18.96	0.26	14.45
200	0.10	4.47	7.46	2.80	7.97	0.76	10.99	0.56	14.08	4.18	4.75	6.33	0.14	18.62	0.29	14.08

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. $k = 5$ is the correctly specified model.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the jackknife variance estimator and normal approximation; **(viii)**: ci‡: average confidence interval length.

Table 3. Bootstrap Inference with Bias Correction, MTE, DGP 1
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$							$\sqrt{n}(\hat{\tau}_{\text{MTE,bc}} - \tau_{\text{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.16	0.14	4.72	4.73	0.05	18.51	0.07	17.59	−0.21	4.93	4.93	0.05	19.31	0.07	18.28
20	0.02	0.63	1.73	4.11	4.46	0.07	16.11	0.07	16.21	0.18	5.26	5.27	0.05	20.63	0.06	19.81
40	0.04	1.26	3.08	3.54	4.69	0.14	13.88	0.12	14.78	1.03	5.11	5.22	0.05	20.05	0.06	19.67
60	0.06	1.90	3.96	3.22	5.11	0.23	12.63	0.20	13.73	1.75	5.02	5.32	0.07	19.68	0.07	19.27
80	0.08	2.53	4.61	3.00	5.50	0.34	11.76	0.28	12.82	2.28	4.91	5.41	0.07	19.24	0.08	18.67
100	0.10	3.16	5.10	2.83	5.83	0.44	11.08	0.38	12.05	2.65	4.78	5.46	0.08	18.72	0.10	18.28
120	0.12	3.79	5.55	2.67	6.16	0.54	10.48	0.48	11.39	2.96	4.66	5.51	0.10	18.25	0.11	17.80
140	0.14	4.43	5.97	2.54	6.49	0.65	9.98	0.59	10.79	3.24	4.57	5.60	0.11	17.90	0.13	17.46
160	0.16	5.06	6.35	2.45	6.81	0.74	9.59	0.69	10.29	3.46	4.43	5.62	0.12	17.36	0.14	17.15
180	0.18	5.69	6.69	2.33	7.09	0.82	9.13	0.77	9.88	3.58	4.35	5.63	0.12	17.04	0.14	16.97
200	0.20	6.32	7.03	2.23	7.38	0.88	8.75	0.84	9.48	3.81	4.22	5.69	0.16	16.56	0.16	16.75

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$							$\sqrt{n}(\hat{\tau}_{\text{MTE,bc}} - \tau_{\text{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.11	0.13	4.85	4.85	0.05	19.00	0.07	17.84	−0.12	4.95	4.95	0.05	19.41	0.07	18.21
20	0.01	0.45	1.42	4.47	4.69	0.06	17.51	0.07	17.03	0.06	5.16	5.16	0.05	20.23	0.06	19.31
40	0.02	0.89	2.73	4.17	4.99	0.10	16.36	0.11	16.17	0.54	5.35	5.38	0.05	20.97	0.06	19.72
60	0.03	1.34	3.78	3.95	5.47	0.16	15.47	0.17	15.38	1.18	5.44	5.57	0.06	21.32	0.07	19.75
80	0.04	1.79	4.62	3.74	5.95	0.24	14.67	0.24	14.73	1.82	5.43	5.73	0.06	21.30	0.09	19.59
100	0.05	2.24	5.27	3.55	6.35	0.32	13.91	0.31	14.09	2.33	5.37	5.86	0.07	21.06	0.10	19.31
120	0.06	2.68	5.77	3.37	6.68	0.41	13.22	0.39	13.59	2.74	5.27	5.94	0.08	20.67	0.10	19.04
140	0.07	3.13	6.27	3.20	7.03	0.51	12.53	0.47	13.12	3.21	5.11	6.04	0.09	20.04	0.12	18.85
160	0.08	3.58	6.67	3.07	7.35	0.59	12.03	0.55	12.72	3.53	5.05	6.16	0.10	19.81	0.13	18.66
180	0.09	4.02	7.07	2.95	7.65	0.68	11.54	0.63	12.30	3.87	4.95	6.28	0.12	19.40	0.15	18.40
200	0.10	4.47	7.42	2.83	7.94	0.74	11.11	0.70	11.91	4.13	4.84	6.36	0.12	18.97	0.15	18.22

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. $k = 5$ is the correctly specified model.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the bootstrap (500 repetitions, Rademacher weights); **(viii)**: ci‡: average confidence interval length.

Table 4. Bootstrap Inference, MTE, DGP 2
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: conventional							$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: percentile ci						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.16	−0.57	6.80	6.82	0.05	26.66	0.00	30.21	−0.96	7.53	7.59	0.06	29.51	0.01	30.21
20	0.02	0.63	−0.41	2.86	2.89	0.06	11.22	0.04	11.38	−1.02	3.16	3.32	0.06	12.39	0.08	11.38
40	0.04	1.26	0.46	2.09	2.14	0.06	8.20	0.05	8.37	−0.37	2.28	2.31	0.06	8.95	0.07	8.37
60	0.06	1.90	1.30	1.91	2.31	0.10	7.48	0.09	7.64	0.28	2.10	2.12	0.05	8.25	0.08	7.64
80	0.08	2.53	1.69	1.87	2.52	0.14	7.32	0.13	7.52	0.43	2.10	2.14	0.06	8.22	0.08	7.52
100	0.10	3.16	2.05	1.85	2.75	0.19	7.23	0.17	7.40	0.60	2.11	2.19	0.06	8.26	0.09	7.40
120	0.12	3.79	2.39	1.82	3.00	0.26	7.14	0.24	7.28	0.79	2.11	2.25	0.06	8.27	0.11	7.28
140	0.14	4.43	2.73	1.80	3.27	0.33	7.06	0.32	7.17	1.01	2.12	2.35	0.07	8.31	0.12	7.17
160	0.16	5.06	3.04	1.77	3.52	0.41	6.94	0.39	7.05	1.23	2.10	2.44	0.09	8.24	0.15	7.05
180	0.18	5.69	3.35	1.74	3.78	0.50	6.82	0.48	6.93	1.48	2.09	2.56	0.10	8.18	0.17	6.93
200	0.20	6.32	3.64	1.72	4.03	0.56	6.75	0.55	6.82	1.74	2.08	2.71	0.12	8.16	0.22	6.82

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: conventional							$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: percentile ci						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.11	−1.39	6.76	6.91	0.06	26.52	0.02	27.91	−1.71	7.06	7.26	0.06	27.67	0.03	27.91
20	0.01	0.45	−1.30	2.99	3.26	0.07	11.72	0.07	11.61	−1.81	3.16	3.64	0.08	12.39	0.10	11.61
40	0.02	0.89	−0.12	2.19	2.19	0.05	8.58	0.05	8.47	−0.79	2.30	2.43	0.06	9.01	0.08	8.47
60	0.03	1.34	0.93	2.02	2.22	0.08	7.91	0.08	7.77	0.10	2.13	2.14	0.05	8.37	0.07	7.77
80	0.04	1.79	1.23	2.00	2.35	0.10	7.83	0.11	7.72	0.17	2.14	2.15	0.05	8.40	0.07	7.72
100	0.05	2.24	1.52	1.98	2.49	0.12	7.74	0.13	7.64	0.25	2.15	2.16	0.05	8.41	0.07	7.64
120	0.06	2.68	1.80	1.97	2.66	0.15	7.70	0.16	7.59	0.34	2.16	2.19	0.05	8.48	0.08	7.59
140	0.07	3.13	2.08	1.95	2.85	0.18	7.64	0.19	7.53	0.44	2.17	2.21	0.06	8.50	0.09	7.53
160	0.08	3.58	2.35	1.94	3.04	0.22	7.60	0.23	7.46	0.55	2.18	2.25	0.06	8.55	0.10	7.46
180	0.09	4.02	2.61	1.92	3.24	0.27	7.54	0.28	7.42	0.68	2.19	2.29	0.06	8.57	0.10	7.42
200	0.10	4.47	2.86	1.91	3.44	0.32	7.48	0.33	7.37	0.80	2.18	2.33	0.07	8.56	0.11	7.37

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. Statistics are centered at the pseudo-true value, 0.545, obtained by using 50 instruments and one million sample size. $k = 50$ is the correctly specified model for estimating the pseudo-true value.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the bootstrap (500 repetitions, Rademacher weights). For the **naive ci**, we first center the bootstrap distribution to suppress its bias correction ability; **(viii)**: ci‡: average confidence interval length.

Table 5. Jackknife Inference, MTE, DGP 2
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.16	−0.97	7.05	7.12	0.06	27.63	0.03	27.97	−1.41	7.73	7.86	0.06	30.31	0.04	27.97
20	0.02	0.63	−0.55	2.85	2.91	0.06	11.18	0.04	12.01	−1.24	3.18	3.41	0.07	12.47	0.08	12.01
40	0.04	1.26	0.42	2.15	2.19	0.05	8.44	0.04	8.78	−0.50	2.41	2.46	0.06	9.45	0.07	8.78
60	0.06	1.90	1.29	1.97	2.35	0.10	7.73	0.08	8.08	0.12	2.24	2.25	0.05	8.80	0.08	8.08
80	0.08	2.53	1.68	1.94	2.56	0.14	7.60	0.11	8.09	0.18	2.31	2.32	0.05	9.06	0.08	8.09
100	0.10	3.16	2.04	1.91	2.80	0.19	7.51	0.14	8.10	0.26	2.37	2.38	0.05	9.28	0.09	8.10
120	0.12	3.79	2.41	1.88	3.05	0.25	7.36	0.19	8.12	0.40	2.37	2.40	0.06	9.29	0.09	8.12
140	0.14	4.43	2.74	1.85	3.31	0.31	7.25	0.24	8.10	0.52	2.43	2.48	0.06	9.51	0.10	8.10
160	0.16	5.06	3.05	1.84	3.56	0.38	7.20	0.30	8.09	0.55	2.46	2.52	0.06	9.66	0.11	8.09
180	0.18	5.69	3.36	1.82	3.82	0.47	7.14	0.36	8.08	0.70	2.58	2.67	0.06	10.10	0.12	8.08
200	0.20	6.32	3.66	1.80	4.08	0.55	7.04	0.42	8.06	0.85	2.61	2.75	0.07	10.23	0.14	8.06

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.11	−1.68	6.91	7.11	0.06	27.09	0.03	27.20	−2.00	7.22	7.49	0.06	28.28	0.05	27.20
20	0.01	0.45	−1.31	2.99	3.26	0.08	11.71	0.07	11.96	−1.84	3.17	3.67	0.09	12.44	0.09	11.96
40	0.02	0.89	−0.08	2.20	2.20	0.05	8.62	0.05	8.71	−0.77	2.33	2.45	0.06	9.13	0.08	8.71
60	0.03	1.34	0.97	2.04	2.26	0.08	8.01	0.08	8.02	0.09	2.18	2.18	0.05	8.54	0.07	8.02
80	0.04	1.79	1.28	2.01	2.39	0.10	7.90	0.09	8.03	0.13	2.19	2.20	0.05	8.59	0.07	8.03
100	0.05	2.24	1.56	1.99	2.53	0.12	7.79	0.11	8.04	0.15	2.21	2.21	0.05	8.65	0.07	8.04
120	0.06	2.68	1.85	1.98	2.71	0.15	7.76	0.13	8.05	0.21	2.25	2.26	0.05	8.83	0.07	8.05
140	0.07	3.13	2.12	1.97	2.90	0.20	7.72	0.17	8.06	0.27	2.27	2.29	0.05	8.90	0.08	8.06
160	0.08	3.58	2.39	1.95	3.08	0.24	7.62	0.20	8.07	0.30	2.30	2.32	0.05	9.00	0.08	8.07
180	0.09	4.02	2.65	1.93	3.28	0.28	7.57	0.24	8.08	0.35	2.31	2.33	0.06	9.05	0.08	8.08
200	0.10	4.47	2.91	1.91	3.48	0.33	7.50	0.28	8.08	0.40	2.33	2.36	0.05	9.12	0.08	8.08

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. Statistics are centered at the pseudo-true value, 0.545, obtained by using 50 instruments and one million sample size. $k = 50$ is the correctly specified model for estimating the pseudo-true value.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the jackknife variance estimator and normal approximation; **(viii)**: ci‡: average confidence interval length.

Table 6. Bootstrap Inference with Bias Correction, MTE, DGP 2
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.16	−1.00	6.89	6.96	0.05	27.02	0.09	24.10	−1.46	7.55	7.69	0.05	29.58	0.09	25.90
20	0.02	0.63	−0.60	2.91	2.97	0.06	11.42	0.06	11.27	−1.30	3.26	3.51	0.07	12.78	0.08	12.54
40	0.04	1.26	0.42	2.12	2.16	0.05	8.32	0.05	8.33	−0.50	2.37	2.42	0.06	9.27	0.05	9.34
60	0.06	1.90	1.25	1.95	2.32	0.10	7.65	0.11	7.59	0.09	2.21	2.21	0.05	8.67	0.05	8.73
80	0.08	2.53	1.65	1.93	2.54	0.15	7.56	0.15	7.48	0.16	2.29	2.30	0.05	8.99	0.05	8.97
100	0.10	3.16	2.01	1.91	2.77	0.19	7.47	0.20	7.36	0.26	2.34	2.36	0.04	9.18	0.05	9.18
120	0.12	3.79	2.35	1.88	3.01	0.24	7.35	0.26	7.26	0.30	2.37	2.39	0.05	9.30	0.05	9.37
140	0.14	4.43	2.72	1.85	3.29	0.31	7.26	0.33	7.14	0.48	2.45	2.49	0.05	9.59	0.06	9.58
160	0.16	5.06	3.04	1.84	3.56	0.39	7.21	0.40	7.05	0.57	2.47	2.54	0.06	9.70	0.06	9.75
180	0.18	5.69	3.34	1.80	3.79	0.47	7.04	0.47	6.94	0.69	2.49	2.58	0.05	9.75	0.06	9.89
200	0.20	6.32	3.62	1.78	4.03	0.54	6.97	0.56	6.83	0.79	2.53	2.65	0.06	9.93	0.06	10.00

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
5	0.00	0.11	−1.82	7.04	7.27	0.05	27.60	0.10	24.83	−2.15	7.36	7.66	0.05	28.84	0.10	25.82
20	0.01	0.45	−1.42	2.99	3.31	0.07	11.72	0.08	11.46	−1.95	3.18	3.73	0.09	12.46	0.10	12.18
40	0.02	0.89	−0.18	2.18	2.19	0.06	8.56	0.06	8.45	−0.87	2.32	2.48	0.07	9.09	0.07	8.99
60	0.03	1.34	0.88	1.98	2.17	0.07	7.77	0.08	7.71	0.00	2.11	2.11	0.05	8.29	0.05	8.32
80	0.04	1.79	1.18	1.97	2.30	0.09	7.72	0.10	7.67	0.01	2.16	2.16	0.06	8.49	0.05	8.48
100	0.05	2.24	1.47	1.96	2.45	0.11	7.69	0.12	7.61	0.04	2.19	2.19	0.05	8.58	0.05	8.63
120	0.06	2.68	1.74	1.93	2.60	0.14	7.58	0.14	7.56	0.08	2.23	2.23	0.05	8.75	0.05	8.78
140	0.07	3.13	2.02	1.92	2.79	0.18	7.54	0.19	7.50	0.11	2.26	2.26	0.05	8.87	0.05	8.89
160	0.08	3.58	2.30	1.90	2.98	0.23	7.44	0.23	7.43	0.16	2.28	2.28	0.05	8.92	0.05	9.06
180	0.09	4.02	2.56	1.88	3.18	0.28	7.37	0.27	7.39	0.26	2.26	2.28	0.05	8.88	0.05	9.16
200	0.10	4.47	2.82	1.87	3.39	0.33	7.34	0.33	7.33	0.28	2.32	2.34	0.05	9.09	0.05	9.30

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. Statistics are centered at the pseudo-true value, 0.545, obtained by using 50 instruments and one million sample size. $k = 50$ is the correctly specified model for estimating the pseudo-true value.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the bootstrap (500 repetitions, Rademacher weights); **(viii)**: ci‡: average confidence interval length.

Table 7. Bootstrap Inference, MTE, DGP 3
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: **conventional**							$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: **percentile ci**						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
6	0.01	0.19	18.10	14.81	23.39	0.21	58.07	0.16	59.52	17.24	15.22	22.99	0.19	59.65	0.15	59.52
11	0.01	0.35	15.55	13.29	20.46	0.19	52.10	0.14	53.43	14.39	14.54	20.45	0.15	56.99	0.14	53.43
21	0.02	0.66	2.23	7.03	7.37	0.06	27.54	0.03	29.03	0.40	8.01	8.02	0.05	31.38	0.04	29.03
26	0.03	0.82	2.86	6.87	7.44	0.07	26.94	0.03	28.48	0.57	8.06	8.08	0.05	31.61	0.04	28.48
56	0.06	1.77	5.70	6.10	8.35	0.14	23.90	0.11	23.95	2.46	8.30	8.65	0.06	32.52	0.11	23.95
61	0.06	1.93	6.15	5.92	8.54	0.16	23.20	0.12	23.51	2.77	8.11	8.57	0.06	31.78	0.12	23.51
126	0.13	3.98	9.26	4.59	10.33	0.51	17.98	0.54	16.47	6.14	6.72	9.10	0.14	26.33	0.34	16.47
131	0.13	4.14	9.53	4.53	10.55	0.54	17.74	0.58	16.25	6.38	6.62	9.19	0.15	25.94	0.36	16.25
252	0.25	7.97	12.64	3.33	13.07	0.98	13.06	0.99	11.30	10.82	4.86	11.86	0.59	19.06	0.86	11.30
257	0.26	8.13	12.79	3.32	13.21	0.98	13.02	0.99	11.21	11.01	4.85	12.03	0.61	19.01	0.86	11.21

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: **conventional**							$\sqrt{n}(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})$: **percentile ci**						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
6	0.00	0.13	24.05	14.68	28.18	0.36	57.56	0.35	57.87	23.46	14.87	27.77	0.34	58.30	0.33	57.87
11	0.01	0.25	20.55	13.34	24.50	0.31	52.30	0.29	52.96	19.69	14.00	24.16	0.27	54.86	0.28	52.96
21	0.01	0.47	1.47	7.07	7.22	0.05	27.73	0.03	28.34	0.12	7.53	7.53	0.05	29.51	0.04	28.34
26	0.01	0.58	1.94	6.99	7.25	0.06	27.39	0.04	28.14	0.18	7.59	7.59	0.06	29.76	0.04	28.14
56	0.03	1.25	4.78	6.77	8.29	0.11	26.56	0.08	26.32	1.62	8.32	8.47	0.06	32.60	0.08	26.32
61	0.03	1.36	5.18	6.75	8.51	0.12	26.47	0.09	26.13	1.81	8.36	8.55	0.06	32.77	0.09	26.13
126	0.06	2.82	8.56	5.99	10.44	0.27	23.46	0.33	21.02	4.41	8.28	9.38	0.08	32.44	0.23	21.02
131	0.07	2.93	8.82	5.93	10.63	0.30	23.25	0.35	20.94	4.59	8.22	9.41	0.09	32.20	0.23	20.94
252	0.13	5.63	12.92	4.50	13.68	0.83	17.63	0.86	15.75	8.89	6.44	10.98	0.27	25.23	0.55	15.75
257	0.13	5.75	13.10	4.47	13.84	0.84	17.54	0.88	15.58	9.03	6.41	11.08	0.27	25.13	0.56	15.58

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Power series expansion is used to estimate nonlinear propensity score. No model is correctly specified, and the misspecification error shrinks as k increases. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the bootstrap (500 repetitions, Rademacher weights). For the **naive ci**, we first center the bootstrap distribution to suppress its bias correction ability; **(viii)**: ci‡: average confidence interval length.

Table 8. Jackknife Inference, MTE, DGP 3
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
6	0.01	0.19	17.82	14.67	23.08	0.22	57.49	0.20	58.47	16.90	15.20	22.73	0.19	59.58	0.19	58.47
11	0.01	0.35	15.31	13.06	20.12	0.20	51.19	0.16	53.94	13.98	14.48	20.13	0.15	56.77	0.17	53.94
21	0.02	0.66	2.06	6.93	7.23	0.06	27.17	0.04	28.35	0.09	8.15	8.15	0.05	31.93	0.04	28.35
26	0.03	0.82	2.71	6.75	7.28	0.07	26.47	0.04	28.40	0.17	8.29	8.29	0.05	32.48	0.05	28.40
56	0.06	1.77	5.78	6.13	8.43	0.14	24.03	0.08	27.92	1.52	9.87	9.99	0.06	38.69	0.10	27.92
61	0.06	1.93	6.24	6.07	8.71	0.16	23.80	0.08	27.67	1.85	9.91	10.08	0.06	38.86	0.10	27.67
126	0.13	3.98	9.31	4.73	10.44	0.49	18.52	0.26	24.00	4.00	9.90	10.67	0.07	38.80	0.21	24.00
131	0.13	4.14	9.57	4.67	10.65	0.53	18.30	0.28	23.77	4.13	9.85	10.68	0.07	38.61	0.22	23.77
252	0.25	7.97	12.61	3.34	13.05	0.97	13.11	0.86	18.29	7.67	8.07	11.13	0.14	31.63	0.44	18.29
257	0.26	8.13	12.75	3.32	13.18	0.97	13.03	0.87	18.19	7.68	8.05	11.12	0.15	31.54	0.44	18.19

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
6	0.00	0.13	24.31	14.22	28.16	0.40	55.75	0.39	56.71	23.77	14.43	27.81	0.37	56.56	0.38	56.71
11	0.01	0.25	20.52	13.00	24.29	0.34	50.95	0.33	52.65	19.66	13.81	24.03	0.29	54.12	0.31	52.65
21	0.01	0.47	1.67	6.98	7.18	0.05	27.37	0.05	27.65	0.31	7.56	7.57	0.05	29.64	0.06	27.65
26	0.01	0.58	2.16	6.90	7.23	0.06	27.04	0.06	27.85	0.35	7.61	7.62	0.05	29.84	0.06	27.85
56	0.03	1.25	4.95	6.47	8.14	0.12	25.36	0.08	28.17	1.33	8.55	8.65	0.05	33.51	0.08	28.17
61	0.03	1.36	5.32	6.36	8.29	0.13	24.95	0.09	28.35	1.42	8.63	8.74	0.05	33.82	0.08	28.35
126	0.06	2.82	8.60	5.52	10.22	0.34	21.63	0.18	27.41	2.80	9.34	9.75	0.07	36.63	0.14	27.41
131	0.07	2.93	8.88	5.50	10.45	0.35	21.57	0.19	27.36	2.93	9.39	9.83	0.07	36.80	0.15	27.36
252	0.13	5.63	12.85	4.38	13.58	0.84	17.17	0.61	23.05	5.67	9.02	10.65	0.09	35.35	0.27	23.05
257	0.13	5.75	13.04	4.37	13.76	0.85	17.11	0.63	22.94	5.80	8.99	10.69	0.10	35.23	0.28	22.94

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Power series expansion is used to estimate nonlinear propensity score. No model is correctly specified, and the misspecification error shrinks as k increases. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the jackknife variance estimator and normal approximation; **(viii)**: ci‡: average confidence interval length.

Table 9. Bootstrap Inference with Bias Correction, MTE, DGP 3
Nominal Level: 0.05

(a) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
6	0.01	0.19	17.92	14.81	23.25	0.22	58.06	0.32	54.47	17.14	15.33	22.99	0.19	60.11	0.28	56.65
11	0.01	0.35	15.55	13.33	20.48	0.20	52.24	0.29	49.13	14.30	15.02	20.74	0.15	58.87	0.22	55.40
21	0.02	0.66	2.12	6.98	7.29	0.06	27.35	0.10	25.46	0.25	8.10	8.11	0.05	31.76	0.08	29.56
26	0.03	0.82	2.75	6.78	7.32	0.07	26.57	0.11	25.00	0.37	8.25	8.26	0.05	32.35	0.07	30.13
56	0.06	1.77	5.58	6.11	8.28	0.14	23.96	0.21	22.92	1.45	9.85	9.96	0.06	38.61	0.07	33.76
61	0.06	1.93	6.02	5.96	8.47	0.16	23.35	0.23	22.76	1.50	9.83	9.94	0.06	38.52	0.07	34.18
126	0.13	3.98	9.13	4.42	10.15	0.53	17.34	0.59	18.09	3.79	9.20	9.95	0.07	36.08	0.09	33.63
131	0.13	4.14	9.42	4.44	10.42	0.56	17.41	0.62	17.88	4.10	9.34	10.20	0.07	36.62	0.10	33.92
252	0.25	7.97	12.54	3.25	12.95	0.97	12.74	0.98	12.01	7.73	7.96	11.09	0.15	31.19	0.26	27.85
257	0.26	8.13	12.68	3.24	13.09	0.97	12.72	0.98	11.82	7.93	7.92	11.21	0.16	31.04	0.26	27.86

(b) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\texttt{MTE,bc}} - \tau_{\texttt{MTE}})$						
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡	bias	sd	$\sqrt{\text{mse}}$	size†	ci†	size‡	ci‡
k																
6	0.00	0.13	24.18	14.56	28.22	0.36	57.08	0.43	54.69	23.55	14.83	27.83	0.34	58.12	0.41	55.89
11	0.01	0.25	20.50	13.37	24.48	0.32	52.39	0.39	49.69	19.57	14.11	24.12	0.27	55.30	0.33	52.89
21	0.01	0.47	1.44	7.13	7.28	0.06	27.96	0.08	26.06	0.10	7.71	7.72	0.05	30.24	0.07	28.33
26	0.01	0.58	1.89	7.00	7.25	0.06	27.44	0.08	25.89	0.14	7.74	7.74	0.05	30.35	0.07	28.84
56	0.03	1.25	4.59	6.70	8.13	0.10	26.28	0.15	24.76	0.86	8.94	8.98	0.05	35.04	0.06	31.89
61	0.03	1.36	4.99	6.64	8.31	0.11	26.03	0.16	24.69	0.87	9.13	9.17	0.06	35.78	0.07	32.31
126	0.06	2.82	8.38	5.70	10.13	0.30	22.34	0.36	22.37	2.33	9.85	10.12	0.06	38.60	0.07	35.27
131	0.07	2.93	8.69	5.65	10.36	0.31	22.15	0.39	22.34	2.60	9.82	10.16	0.06	38.51	0.07	35.45
252	0.13	5.63	12.72	4.39	13.46	0.82	17.22	0.85	17.14	5.51	8.91	10.47	0.09	34.93	0.14	33.18
257	0.13	5.75	12.91	4.39	13.63	0.84	17.20	0.87	17.03	5.62	9.03	10.63	0.10	35.38	0.13	33.14

Notes. The marginal treatment effect is evaluated at $a = 0.5$, or equivalently it is $\hat{\theta}_2 + \hat{\theta}_3$. Power series expansion is used to estimate nonlinear propensity score. No model is correctly specified, and the misspecification error shrinks as k increases. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively.

(i) k: number of instruments used for propensity score estimation; **(ii)** bias: empirical bias; **(iii)** sd: empirical standard deviation; **(iv)** mse: empirical mean squared error (i.e. bias^2+sd^2); **(v)** size†: empirical size of the level-0.05 test, where the t-statistic is constructed with the (infeasible) oracle standard deviation; **(vi)** ci†: average confidence interval length of the t-test using the (infeasible) oracle standard deviation; **(vii)** size‡: empirical size of the level-0.05 test based on the bootstrap (500 repetitions, Rademacher weights); **(viii)**: ci‡: average confidence interval length.

Table 10. Summary Statistics ($n = 1,747$)

Variable	Description	college= 0 ($n = 882$)	college= 1 ($n = 865$)
wage91	log wage in 1991	2.209	2.550
		[0.441]	[0.496]
college	college attendance	0.000	1.000
cAFQT	corrected AFQT score	−0.045	0.952
		[0.867]	[0.750]
exp	working experience	10.100	6.840
		[3.126]	[3.252]
YoB57	1=born in 1957	0.098	0.103
YoB58	1=born in 1958	0.083	0.112
YoB59	1=born in 1959	0.120	0.089
YoB60	1=born in 1960	0.137	0.125
YoB61	1=born in 1961	0.127	0.133
YoB62	1=born in 1962	0.167	0.169
YoB63	1=born in 1963	0.136	0.141
urban14	1=urban residency at 14	0.700	0.790
eduMom	mom education	11.310	12.910
		[2.106]	[2.279]
numSiblings	number of siblings	3.263	2.585
		[2.084]	[1.645]
pub4	1= presence of public 4 year college in county of residence at 14	0.463	0.588
avgTui17	average tuition in public 4 year colleges in county of residence at 17	22.020	21.110
		[7.873]	[8.068]
avgUne17Perm	average permanent unemployment in state of residence at 17	6.294	6.208
		[1.016]	[0.954]
avgWag17Perm	log average permanent wage in county of residence at 17	10.270	10.300
		[0.180]	[0.195]
avgUne17	average unemployment in state of residence at 17	7.080	7.085
		[1.785]	[1.845]
avgWag17	log average wage in county of residence at 17	10.280	10.270
		[0.162]	[0.165]
avgUne91	average unemployment in state of residence in 1991	6.797	6.823
		[1.331]	[1.198]
avgWag91	log average wage in county of residence in 1991	10.260	10.320
		[0.160]	[0.166]

Notes. Standard deviations in square brackets.

Table 11. Marginal Treatment Effects ($p = 2$)

	$\hat{\tau}_{\texttt{MTE}}(0.2)$					$\hat{\tau}_{\texttt{MTE}}(0.5)$					$\hat{\tau}_{\texttt{MTE}}(0.8)$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
no bias correction	0.418	0.401	0.324	0.307	0.305	0.072	0.059	0.110	0.097	0.069	−0.274	−0.283	−0.104	−0.113	−0.167
	(0.107)	(0.097)	(0.084)	(0.082)	(0.089)	(0.052)	(0.052)	(0.045)	(0.042)	(0.035)	(0.159)	(0.141)	(0.116)	(0.108)	(0.107)
bias corrected	0.460	0.523	0.414	0.422	0.362	0.094	0.057	0.102	0.090	0.072	−0.273	−0.410	−0.210	−0.241	−0.218
	(0.161)	(0.145)	(0.132)	(0.132)	(0.121)	(0.067)	(0.074)	(0.056)	(0.055)	(0.044)	(0.232)	(0.214)	(0.174)	(0.175)	(0.143)
Outcome Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[b]exp (and squared)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[c]avgUne91, avgWag91	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
k								47							
$k/\sqrt{n}$								1.12							
Selection Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[d]instruments	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[e]× cAFQT, eduMom, numSib	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[f]linear interactions		✓		✓			✓		✓			✓		✓	
[g]cohort interactions			✓	✓				✓	✓				✓	✓	
[h]logit					✓					✓					✓
k	35	45	56	66	35	35	45	56	66	35	35	45	56	66	35
$k/\sqrt{n}$	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84

Notes. The marginal treatment effects are estimated at 0.2, 0.5 and 0.8, and are evaluated at mean values of the covariates. The estimated propensity score enters quadratically. Bias correction is based on the jackknife method, and standard error are obtained by inverting the 95% bootstrap confidence interval (500 bootstrap repetitions).
a. Linear and square terms of corrected AFQT score, education of mom, number of siblings, average permanent local unemployment rate and wage rate at age 17; urban residency at age 14; and cohort dummies.
b. Experience in 1991, and squared.
c. Average local unemployment and wage rate in 1991.
d. Raw instruments, including presence of four year college at age 14, average local college tuition at age 17, average local unemployment rate and wage rate at age 17.
e. Interaction of the the raw instruments with corrected AFQT score, education of mom, and number of siblings.
f. First order interactions among corrected AFQT score, education of mom, number of siblings, average permanent local unemployment rate and wage rate at age 17.
g. Interactions of cohort dummies with corrected AFQT score, education of mom and number of siblings.
h. Logit model is used to estimate the selection equation.

Table 12. Marginal Treatment Effects ($p = 3$)

	$\hat{\tau}_{\texttt{MTE}}(0.2)$					$\hat{\tau}_{\texttt{MTE}}(0.5)$					$\hat{\tau}_{\texttt{MTE}}(0.8)$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
no bias correction	0.430	0.414	0.317	0.291	0.338	0.062	0.050	0.117	0.110	0.053	−0.267	−0.277	−0.112	−0.127	−0.130
	(0.118)	(0.101)	(0.083)	(0.086)	(0.095)	(0.061)	(0.064)	(0.046)	(0.047)	(0.041)	(0.176)	(0.141)	(0.120)	(0.120)	(0.116)
bias corrected	0.483	0.561	0.384	0.391	0.412	0.074	0.028	0.128	0.113	0.049	−0.269	−0.398	−0.233	−0.264	−0.161
	(0.175)	(0.160)	(0.134)	(0.146)	(0.130)	(0.076)	(0.086)	(0.062)	(0.067)	(0.052)	(0.253)	(0.220)	(0.183)	(0.200)	(0.150)
Outcome Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[b]exp (and squared)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[c]avgUne91, avgWag91	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
k								48							
$k/\sqrt{n}$								1.15							
Selection Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[d]instruments	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[e]× cAFQT, eduMom, numSib	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[f]linear interactions		✓		✓			✓		✓			✓		✓	
[g]cohort interactions			✓	✓				✓	✓				✓	✓	
[h]logit					✓					✓					✓
k	35	45	56	66	35	35	45	56	66	35	35	45	56	66	35
$k/\sqrt{n}$	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84

Notes. The marginal treatment effects are estimated at 0.2, 0.5 and 0.8, and are evaluated at mean value of the covariates. The estimated propensity score enters up to third order. Bias correction is based on the jackknife method, and standard error are obtained by inverting the 95% bootstrap confidence interval (500 bootstrap repetitions).

Table 13. Marginal Treatment Effects ($p = 4$)

	$\hat{\tau}_{\texttt{MTE}}(0.2)$					$\hat{\tau}_{\texttt{MTE}}(0.5)$					$\hat{\tau}_{\texttt{MTE}}(0.8)$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
no bias correction	0.433	0.416	0.329	0.301	0.358	0.069	0.055	0.121	0.114	0.044	−0.267	−0.278	−0.124	−0.137	−0.170
	(0.113)	(0.100)	(0.087)	(0.085)	(0.105)	(0.061)	(0.063)	(0.048)	(0.049)	(0.043)	(0.154)	(0.146)	(0.131)	(0.113)	(0.143)
bias corrected	0.485	0.558	0.407	0.400	0.441	0.086	0.033	0.130	0.113	0.034	−0.282	−0.392	−0.257	−0.271	−0.227
	(0.161)	(0.159)	(0.126)	(0.120)	(0.148)	(0.085)	(0.092)	(0.060)	(0.071)	(0.056)	(0.225)	(0.232)	(0.178)	(0.180)	(0.200)
Outcome Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[b]exp (and squared)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[c]avgUne91, avgWag91	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
k								49							
$k/\sqrt{n}$								1.17							
Selection Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[d]instruments	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[e]× cAFQT, eduMom, numSib	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[f]linear interactions		✓		✓			✓		✓			✓		✓	
[g]cohort interactions			✓	✓				✓	✓				✓	✓	
[h]logit					✓					✓					✓
k	35	45	56	66	35	35	45	56	66	35	35	45	56	66	35
$k/\sqrt{n}$	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84

Notes. The marginal treatment effects are estimated at 0.2, 0.5 and 0.8, and are evaluated at mean value of the covariates. The estimated propensity score enters up to fourth order. Bias correction is based on the jackknife method, and standard error are obtained by inverting the 95% bootstrap confidence interval (500 bootstrap repetitions).

Table 14. Marginal Treatment Effects ($p = 5$)

	$\hat{\tau}_{\texttt{MTE}}(0.2)$					$\hat{\tau}_{\texttt{MTE}}(0.5)$					$\hat{\tau}_{\texttt{MTE}}(0.8)$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
no bias correction	0.414	0.393	0.285	0.261	0.323	0.086	0.073	0.155	0.145	0.079	−0.285	−0.300	−0.159	−0.170	−0.188
	(0.108)	(0.110)	(0.104)	(0.101)	(0.112)	(0.066)	(0.066)	(0.062)	(0.056)	(0.054)	(0.182)	(0.153)	(0.120)	(0.116)	(0.139)
bias corrected	0.486	0.556	0.489	0.469	0.415	0.089	0.040	0.061	0.063	0.061	−0.301	−0.411	−0.197	−0.205	−0.237
	(0.160)	(0.174)	(0.189)	(0.185)	(0.153)	(0.109)	(0.124)	(0.130)	(0.131)	(0.070)	(0.246)	(0.223)	(0.181)	(0.188)	(0.185)
Outcome Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[b]exp (and squared)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[c]avgUne91, avgWag91	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
k								50							
$k/\sqrt{n}$								1.20							
Selection Eqn.															
[a]baseline	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[d]instruments	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[e]× cAFQT, eduMom, numSib	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[f]linear interactions		✓		✓			✓		✓			✓		✓	
[g]cohort interactions			✓	✓				✓	✓				✓	✓	
[h]logit					✓					✓					✓
k	35	45	56	66	35	35	45	56	66	35	35	45	56	66	35
$k/\sqrt{n}$	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84	0.84	1.08	1.34	1.58	0.84

Notes. The marginal treatment effects are estimated at 0.2, 0.5 and 0.8, and are evaluated at mean value of the covariates. The estimated propensity score enters up to fifth order. Bias correction is based on the jackknife method, and standard error are obtained by inverting the 95% bootstrap confidence interval (500 bootstrap repetitions).