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

Matias D. Cattaneo† Michael Jansson‡ Xinwei Ma§

October 13, 2017

Abstract

We study the implications of including many covariates in a first-step estimate entering a two-step estimation procedure. We find that a first order bias emerges when the number of *included* covariates is "large" relative to the square-root of sample size, rendering standard inference procedures invalid. We show that the jackknife is able to estimate this "many covariates" bias consistently, thereby delivering a new automatic bias-corrected two-step point estimator. The jackknife also consistently estimates the standard error of the original two-step point estimator. For inference, we develop a valid post-bias-correction bootstrap approximation that accounts for the additional variability introduced by the jackknife bias-correction. We find that the jackknife bias-corrected point estimator and the bootstrap post-bias-correction inference perform excellent in simulations, offering important improvements over conventional two-step point estimators and inference procedures, which are not robust to including many covariates. We apply our results to an array of distinct treatment effect, policy evaluation and other applied microeconomics settings. In particular, we discuss production function and marginal treatment effect estimation in detail.

*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 and Department of Statistics, University of Michigan.

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

§Department of Economics and Department of Statistics, University of Michigan.

1 Introduction

Two-step estimators are very important and widely used in empirical work in Economics and other disciplines. This approach involves two estimation steps: first an unknown quantity is estimated, and then this estimate is plugged in a moment condition to form the second and final point estimator of interest. For example, inverse probability weighting (IPW) and generated regressors methods fit naturally into this framework, both used routinely in treatment effect and policy evaluation settings. In practice, researchers often include many covariates in the first-step estimation procedure in an attempt to flexibly control for as many confounders as possible, even after model selection or model shrinking has been used to select out some of all available covariates. Conventional (post-model selection) estimation and inference results in this context, however, assume that the number of covariates included in the estimation is "small" relative to the sample size, and hence the effect of overfitting in the first estimation step is ignored in current practice.

We show that two-step estimators can be severely biased when too many covariates are included in a linear-in-parameters first-step, a fact that leads to invalid inference procedures even in large samples. This crucial, but often overlooked fact implies that many empirical conclusions will be incorrect whenever many covariates are used. For example, we find from a very simple simulation setup with a first step estimated with 80 i.i.d. variables, sample size of $2,000$, and even no misspecification bias, that a conventional 95% confidence interval covers the true parameter with probability 76% due to the presence of the many covariates bias we highlight in this paper (Table 1 below).[1] This striking result is not specific to our simulation setting, as our general results apply broadly to many other settings in treatment effect, policy evaluation and applied microeconomics: IPW estimation under unconfoundedness, semiparametric difference-in-differences, local average response function estimation, marginal treatment effects, control function methods, and production function estimation, just to mention a few other popular examples.

We illustrate the usefulness of our results by considering several applications in applied microeconomics. In particular, we discuss in detail production function (Olley and Pakes, 1996) and marginal treatment effect (Heckman and Vytlacil, 2005) estimation when possibly many covari-

[1] Including 80 regressors is quite common in empirical work: e.g., settings with 50 residential dummy indicators, a few covariates entering linearly and quadratically, and perhaps some interactions among these variables. Section SA-9 of the Supplemental Appendix collects a sample of empirical papers employing two-step estimators with arguably many covariates.

ates/instruments are present. The latter application offers new estimation and inference results in instrumental variable (IV) settings allowing for treatment effect heterogeneity and many covariates/instruments.

The presence of the generic many covariates bias we highlight implies that developing more robust procedures accounting for possibly many covariates entering the first step estimation is highly desirable. Such robust methods would give more credible empirical results, thereby providing more plausible testing of substantive hypotheses as well as more reliable policy prescriptions. With this goal in mind, we show that jackknife bias-correction is able to remove the many covariates bias we uncover in a fully automatic way. Under mild conditions on the design matrix, we prove consistency of the jackknife bias and variance estimators, even when many covariates are included in the first-step estimation. Indeed, our simulations in the context of MTE estimation show that jackknife bias-correction is quite effective in removing the many covariates bias, exhibiting roughly a 50% bias reduction (Table 1 below). We also show that the mean squared error of the jackknife bias-corrected estimator is substantially reduced whenever many instruments are included. More generally, our results give a new, fully automatic, jackknife bias-corrected two-step estimator with demonstrably superior properties to use in applications.

For inference, while jackknife bias correction and variance estimation delivers a valid Gaussian distributional approximation in large samples, we find in our simulations that the associated inference procedures do not perform as well in small samples. As discussed in Calonico, Cattaneo, and Farrell (2017) in the context of kernel-based nonparametric inference, a crucial underlying issue is that bias correction introduces additional variability not accounted for in samples of moderate size (we confirm this finding in our simulations). Therefore, to develop better inference procedures in finite samples, we also establish validity of a bootstrap method applied to the jackknife-based bias-corrected Studentized statistic, which can be used to construct valid confidence intervals and conduct valid hypothesis tests in a fully automatic way. This procedure is a hybrid of the wild bootstrap (first-step estimation) and the multiplier bootstrap (second-step estimation), which is fast and easy to implement in practice because it avoids recomputing the relatively high-dimensional portion of the first estimation step. Under generic regularity conditions, we show that this bootstrap procedure successfully approximates the finite sample distribution of the bias-corrected jackknife-based Studentized statistic, a result that is also borne out in our simulation study.

Put together, our results not only highlight the important negative implications of overfitting the first-step estimate in generic two-step estimation problems, which leads to a first order many covariates bias in the distributional approximation, but also provide fully automatic resampling methods to construct more robust estimators and inference procedures. Furthermore, because our results remain asymptotically valid when only a few covariates are used, they provide strict asymptotic improvement over conventional methods currently used in practice. All our results are fully automatic and do not require additional knowledge about the data generating process, which implies that they can be easily implemented in empirical work using straightforward resampling methods in any computing platform.

In the remainder of this introduction section we discuss some of the many literatures this paper is connected to. Then, the rest of the paper unfolds as follows. Section 2 introduces the setup and gives an overview of our results. Section 3 gives details on the main properties of the two-step estimator, in particular characterizing the non-vanishing bias due to many covariates entering the first-step estimate. Section 4 establishes validity of the jackknife bias and variance estimator, and therefore presents our proposed bias-corrected two-step estimator, while Section 5 establishes valid distributional approximations for the jackknife-based bias-corrected Studentized statistic using a carefully modified bootstrap method. Section 6 applies our main results to two examples: production function (Olley and Pakes, 1996) and marginal treatment effect (Heckman and Vytlacil, 2005) estimation, while six other treatment effect, program evaluation and applied microeconomics examples are discussed in the Supplemental Appendix (Section SA-5) to conserve space. Section 7 summarizes the main results from an extensive Monte Carlo experiment and an empirical illustration using real data, building on the work of Carneiro, Heckman, and Vytlacil (2011). Finally, Section 8 concludes. The long Supplemental Appendix also contains methodological and technical details, includes the theoretical proofs of our main theorems as well as other related results, and reports additional numerical evidence.

1.1 Related Literature

Our work is related to several interconnected literatures in econometrics and statistics.

Two-step Semiparametrics. From a classical semiparametric perspective, when the many covariates included in the first-step estimate are taken as basis expansions of some underlying fixed

dimension regressor, our final estimator becomes a two-step semiparametric estimator with a nonparametric series-based preliminary estimate. Conventional large sample approximations in this case are well known (e.g., Newey and McFadden, 1994; Chen, 2007; Ichimura and Todd, 2007, and references therein). From this perspective, our paper contributes not only to this classical semiparametric literature, but also to the more recent work in the area, which has developed distributional approximations that are more robust to tuning parameters choices and underlying assumptions (e.g., smoothness). In particular, first, Cattaneo, Crump, and Jansson (2013) and Cattaneo and Jansson (2017) develop approximations for two-step non-linear kernel-based semiparametric estimators when possibly a "small" bandwidth is used, which leads to a first-order bias due to undersmoothing the preliminary kernel-based nonparametric estimate, and show that inference based on the nonparametric bootstrap automatically accounts for the small bandwidth bias explicitly, thereby offering more robust inference procedures in that context.[2] Second, Chernozhukov, Escanciano, Ichimura, and Newey (2017) study the complementary issue of "large" bandwidth or "small" number of series terms, and develop more robust inference procedures in that case. Their approach is to modify the estimating equation so that the resulting new two-step estimator is less sensitive to oversmoothing (i.e., underfitting) the first-step nonparametric estimator. Our paper complements this recent work in semiparametrics by offering new inference procedures for two-step non-linear semiparametric estimators with demonstrably more robust properties to undersmoothing (i.e., overfitting) a first step series-based estimator, results that are not currently available in the semiparametrics literature. See Section 3 for more details.

High-dimensional Models. Our results go beyond semiparametrics because we do not assume (but allow for) the first-step estimate to be a nonparametric series-based estimator. In fact, we do not rely on any specific structure of the covariates in the first step, nor do we rely on asymptotic linear representations. Thus, our results also contribute to the literature on high-dimensional models in statistics and econometrics (e.g., Mammen, 1989, 1993; El Karoui, Bean, Bickel, Lim, and Yu, 2013; Cattaneo, Jansson, and Newey, 2018; Li and Müller, 2017, and references therein) by developing generic distributional approximations for two-step estimators where the first-step

[2] A certain class of *linear* semiparametric estimators has a very different behavior when undersmoothing the first step nonparametric estimator; see Cattaneo, Jansson, and Newey (2017) for discussion and references. In particular, the results in that paper show that undersmoothing leads to an additional variance contribution (due to the underlying linearity of the model), while in the present paper we find a bias contribution instead (due to the non-linearity of the models considered).

estimator is possibly high-dimensional. See also Fan, Lv, and Qi (2011) for a survey and discussion on high-dimensional and ultra-high-dimensional models.[3] A key distinction here is that the class of estimators we consider is defined through a moment condition that is non-linear in the first step estimate (e.g., propensity score, generated regressor, etc.). Previous work on high-dimensional models has focused exclusively on either linear least squares regression or one-step (possibly non-linear) least squares regression. In contrast, this paper covers a large class of two-step non-linear procedures, going well beyond least squares regression for the second step estimation procedure. Most interestingly, our results show formally that when many covariates are included in a first-step estimation the resulting two-step estimator exhibits a bias of order $k/\sqrt{n}$ in the distributional approximation, where k denotes the number of included covariates and n denotes the sample size. This finding contrasts sharply with previous results for high-dimensional linear regression models with many covariates, where it has been found that including many covariates leads to a variance contribution (not a bias contribution as we find herein) in the distributional approximation, which is of order k/n (not $k/\sqrt{n}$ as we find herein). By implication, the many covariates bias we uncover in this paper will have a first-order effect on inference when fewer covariates are included relative to the case of high-dimensional linear regression models.

Ultra-high-dimensional Models and Covariate Selection. Our results also have implications for the recent and rapidly growing literature on inference after covariate/model selection in ultra-high-dimensional settings under sparsity conditions (e.g., Belloni, Chernozhukov, and Hansen, 2014; Farrell, 2015; Belloni, Chernozhukov, Fernández-Val, and Hansen, 2017, and references therein). In this literature, the total number of available covariates/instruments is allowed to be much larger than the sample size, but the final number of *included* covariates/instruments is much smaller than the sample size, as most available covariates are selected out by some penalization or model selection method (e.g., LASSO) employing some form of an sparsity assumption. This implies that the number of included covariates/instruments effectively used for estimation and inference (k in our notation) is much smaller than the sample size, as the underlying distribution theory in that literature requires $k/\sqrt{n} = o(1)$. Therefore, because $k/\sqrt{n} = O(1)$ is the only restriction assumed in this paper, our results shed new light on situations where the number of selected or

[3] We call models high-dimensional when the number of available covariates is at most a fraction of the sample size and ultra-high-dimensional when the number available covariates is larger than the sample size.

included covariates, possibly after model selection, is not "small" relative to the sample size. We formally show that valid inference post-model selection requires that a relatively small number of covariates enter the final specification, since otherwise a first order bias will be present in the distributional approximations commonly employed in practice, thereby invalidating the associated inference procedures. Our results do not employ any sparsity assumption and allow for any kind of regressors, including many fixed effects, provided the first-step estimate can be computed.

Large-(N,T) Panel Data. Our findings are also qualitatively connected to the literature on non-linear panel models with fixed effects (Alvarez and Arellano, 2003; Hahn and Newey, 2004) in at least two ways. First, in that context a first-order bias arises when the number of time periods (T) is proportional to the number of entities (N), just like we uncover a first-order bias when $k \propto \sqrt{n}$, and in both cases this bias can be heuristically attributed to an incidental parameters/overfitting problem. Second, in that literature jackknife bias correction was shown to be able to remove the large-(N,T) bias, just like we establish a similar result in this paper for a class of two-step estimators with high-dimensional first-step. Beyond these two superficial connections, however, our findings are both technically and conceptually quite different from the results already available in the large-(N,T) non-linear panel models with fixed effects literature.

Applications. From a practical perspective, our results offer new inference results for many popular estimators in program evaluation and treatment effect settings (e.g., Abadie and Cattaneo, 2018, and references therein), as well as other areas in empirical microeconomics (e.g., Ackerberg, Benkard, Berry, and Pakes, 2007, and references therein). Section 6.1 discusses production function estimation, which provides new econometric methodology in the context of an IO application, while Section 6.2 considers marginal treatment effect estimation, where we develop new estimation and inference methods in the presence of many covariates/instruments and heterogeneous IV treatment effects. Furthermore, because our results apply to non-linear settings in general, we cover many other settings of interest: (i) IPW under unconfoundedness (e.g., Hirano, Imbens, and Ridder, 2003; Chen, Hong, and Tarozzi, 2008; Cattaneo, 2010, and references therein), (ii) Semiparametric Difference-in-Differences (Abadie, 2005), (iii) Local Average Response Function (Abadie, 2003), (iv) Two-Stage Least Squares and Conditional Moment Restrictions (e.g., Wooldridge, 2010, for a textbook review), and (v) Control Function Methods (e.g., Wooldridge, 2015, and references therein). All these other examples are analyzed in Section SA-5 of the Supplemental Appendix.

2 Setup and Overview of Results

We consider a two-step GMM setting where $\mathbf{w}_i = (\mathbf{y}_i^\mathsf{T}, r_i, \mathbf{z}_i^\mathsf{T})^\mathsf{T}$, $i = 1, 2, \ldots, n$, denotes an observed random sample, and the finite dimensional parameter of interest $\boldsymbol{\theta}_0$ solves uniquely the (possibly over-identified) vector-valued moment condition $\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)] = \mathbf{0}$ with $\mu_i = \mathbb{E}[r_i|\mathbf{z}_i]$. Thus, we specialize the general two-step GMM approach in that we view the unknown scalar μ_i as a "generated regressor" depending on possibly "many" covariates $\mathbf{z}_i \in \mathbb{R}^k$, which we take as the included variables entering the first-step specification (i.e., after model/covariate selection). Our results extend immediately to vector-valued unknown $\boldsymbol{\mu}_i$, albeit with cumbersome notation, as shown in Section SA-4.1 of the Supplemental Appendix. See Section 6.1 for an application with a bivariate first-step.

Given a first-step estimate $\hat{\mu}_i$ of μ_i, which we construct using least-squares projections on the possibly high-dimensional covariate $\mathbf{z}_i$ as discussed further below, we study the two-step estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left| \boldsymbol{\Omega}_n^{1/2} \sum_{i=1}^{n} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}) \right|, \tag{1}$$

where $|\cdot|$ denotes the Euclidean norm, $\Theta \subseteq \mathbb{R}^d$ is the parameter space, and $\boldsymbol{\Omega}_n$ is a (possibly random) positive semi-definite conformable weighting matrix with positive definite probability limit $\boldsymbol{\Omega}_0$. Precise regularity conditions on the known moment function $\mathbf{m}(\cdot)$ are given in Appendix A.

When the dimension of the included variables $\mathbf{z}_i$ is "small" relative to the sample size, $k = o(\sqrt{n})$, textbook large sample theory is valid, and hence estimation and inference can be conducted in the usual way (e.g., Newey and McFadden, 1994). However, when the dimension of the included covariates used to approximate the unknown component μ_i is "large" relative to the sample size, $k = O(\sqrt{n})$, standard distribution theory fails. To be more specific, under fairly general regularity conditions, we show in Section 3 that:

$$\mathscr{V}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \mathscr{B}) \rightsquigarrow \text{Normal}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\rightsquigarrow$ denotes convergence in distribution, with limits always taken as $n \to \infty$ and $k = O(\sqrt{n})$, and $\mathscr{V}$ and $\mathscr{B}$ denoting, respectively, the approximate variance and bias of the estimator $\hat{\boldsymbol{\theta}}$. This result has a key distinctive feature relative to classical textbook results: a first-order bias $\mathscr{B}$ emerges

whenever "many" covariates are included, that is, whenever k is "large" relatively to n in the sense that $k/\sqrt{n} \not\to 0$. A crucial practical implication of this finding is that conventional inference procedures that disregard the presence of the first-order bias will be incorrect even asymptotically, since $\mathscr{V}^{-1/2}\mathscr{B} = O_{\mathbb{P}}(k/\sqrt{n})$ is non-negligible. For example, non-linear treatment effect, instrumental variables and control function estimators employing "many" included covariates in a first-step estimation will be biased, thereby giving over-rejection of the null hypothesis of interest. In Section 7 we illustrate this problem using simulated data in the context of instrumental variable models with many instruments/covariates, where we find that typical hypothesis tests over-reject the null hypothesis four times as often as they should in practically relevant situations.

Putting aside the bias issue when many covariates are used in the first-step estimation, another important issue regarding (2) is the characterization and estimation of the variance $\mathscr{V}$. Because the possibly high-dimensional covariates $\mathbf{z}_i$ are not necessarily assumed to be a series expansion, or other type of convergent sequence of covariates, the variance $\mathscr{V}$ is harder to characterize and estimate. In fact, our distributional approximation leading to (2) is based on a quadratic approximation of $\hat{\boldsymbol{\theta}}$, as opposed to the traditional linear approximation commonly encountered in the semiparametrics literature (Newey, 1994; Chen, 2007; Hahn and Ridder, 2013), thereby giving a more general characterization of the variability of $\hat{\boldsymbol{\theta}}$ with potentially better finite sample properties.

Nevertheless, our first main result (2) suggests that valid inference in two-step GMM settings is possible even when many covariates are included in the first-step estimation, if consistent variance and bias estimators are available. Our second main result (in Section 4) shows that the jackknife offers an easy-to-implement and automatic way to approximate both the variance and the bias:

$$\mathscr{T} \overset{\text{def}}{=} \hat{\mathscr{V}}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \hat{\mathscr{B}}) \rightsquigarrow \text{Normal}(\mathbf{0}, \mathbf{I}), \tag{3}$$

$$\hat{\mathscr{B}} = (n-1)(\hat{\boldsymbol{\theta}}^{(\cdot)} - \hat{\boldsymbol{\theta}}), \qquad \hat{\mathscr{V}} = \frac{n-1}{n}\sum_{\ell=1}^{n}(\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\cdot)})(\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\cdot)})^{\mathsf{T}}, \qquad \hat{\boldsymbol{\theta}}^{(\cdot)} = \frac{1}{n}\sum_{\ell=1}^{n}\hat{\boldsymbol{\theta}}^{(\ell)}, \tag{4}$$

where to construct $\hat{\boldsymbol{\theta}}^{(\ell)}$, the ℓth observation is deleted and then both steps are re-estimated using the remaining observations. Simulation evidence reported in Section 7 confirms that the jackknife provides an automatic data-driven method able to approximate quite well both the bias and the

variance of the estimator $\hat{\boldsymbol{\theta}}$, even when many covariates are included in the first-step estimation procedure. An important virtue of the jackknife is that it can be implemented very fast in special settings, which is particularly important in high-dimensional situations. Indeed, our first-step estimator will be constructed using least-squares approximations, a method that is particularly amenable to jackknifing.

While result (3) could be used for inference in large samples, a potential drawback is that the jackknife bias-correction introduces additional variability not accounted for in samples of moderate size. Therefore, to improve inference further in applications, we develop a new, specifically tailored bootstrap-based distributional approximation to the jackknife-based bias-corrected and Studentized statistic. Our method combines the wild bootstrap (first-step) and the multiplier bootstrap (second-step), while explicitly taking into account the effect of jackknifing under the multiplier bootstrap law (see Section 5 for more details). To be more specific, our third and final main result is:

$$\sup_{t\in\mathbb{R}^d}\Big|\mathbb{P}[\mathscr{T}\leq t]-\mathbb{P}^\star[\mathscr{T}^\star\leq t]\Big|\to_{\mathbb{P}}0,\qquad \mathscr{T}^\star\ \overset{\text{def}}{=}\ \hat{\mathscr{V}}^{\star^{-1/2}}(\hat{\boldsymbol{\theta}}^\star-\hat{\boldsymbol{\theta}}-\hat{\mathscr{B}}^\star),\tag{5}$$

where $\hat{\boldsymbol{\theta}}^\star$ is a bootstrap counterpart of $\hat{\boldsymbol{\theta}}$, $\hat{\mathscr{B}}^\star$ and $\hat{\mathscr{V}}^\star$ are properly weighted jackknife bias and variance estimators under the bootstrap distribution, respectively, and $\mathbb{P}^\star$ is the bootstrap probability law conditional on the data. Our bootstrap approach is fully automatic and captures explicitly the distributional effects of estimating the bias (and variance) using the jackknife, and hence delivers a better finite sample approximation. Simulation evidence reported in Section 7 supports this result.

In sum, valid and more robust inference in two-step GMM settings with possible many covariates entering the first-step estimate can be conducted by combining results (3) and (5). Specifically, our approach requires three simple and automatic stages: (i) constructing the two-step estimator ($\hat{\boldsymbol{\theta}}$), (ii) constructing the jackknife bias and variance estimators ($\hat{\mathscr{B}}$ and $\hat{\mathscr{V}}$), and finally (iii) conducting inference as usual but employing bootstrap quantiles obtained from (5) instead of those from the normal approximation. In the remainder of this paper we formalize these results and illustrate them using simulated as well as real data.

3 The Effect of Including Many Covariates

In this section we formalize the implications of overfitting the first-step estimate entering (1), and show that under fairly general regularity conditions the estimator $\hat{\boldsymbol{\theta}}$, and transformations thereof, exhibit a first-order bias whenever k is "large", that is, whenever $k \propto \sqrt{n}$. The results in this section justify, in particular, the distributional approximation in (2).

3.1 First-Step Estimation

We are interested in understanding the effects of introducing possibly many covariates $\mathbf{z}_i$, that is, in cases where its dimension k is possibly "large" relative to the sample size. For tractability and simplicity, we consider linear approximations to the unknown component:

$$\mu_i = \mathbb{E}[r_i|\mathbf{z}_i] = \mathbf{z}_i^{\mathsf{T}}\boldsymbol{\beta} + \eta_i, \qquad \mathbb{E}[\mathbf{z}_i\eta_i] = \mathbf{0}, \tag{6}$$

where η_i represents the error in the best linear approximation. This motivates the commonly used linear least-squares first-step estimate

$$\hat{\mu}_i = \mathbf{z}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (r_i - \mathbf{z}_i^{\mathsf{T}}\boldsymbol{\beta})^2. \tag{7}$$

This approach is quite common in empirical work. It is possible to allow for non-linear models, but such methods are harder to handle mathematically and usually do not perform well numerically when $\mathbf{z}_i$ is of large dimension. Furthermore, a non-linear approach will be computationally more difficult, as we discuss in more detail below. As shown in the already lengthy Supplemental Appendix (Section SA-9), our proofs explicitly exploit the linear regression representation of $\hat{\mu}_i$ to scale down the already quite involved technical work. Nevertheless, we also conducted preliminary theoretical work to verify that the main results presented below carry over to non-linear least-squares estimators (e.g., logistic regression when r_i is binary).

Using the first-step estimate $\hat{\mu}_i$ in (7), we investigate the implications of introducing possibly many covariates/instruments $\mathbf{z}_i$, and thus our approximations allow for (but do not require that) k being "large" relative to the sample size. In this case conventional inference procedures become invalid due to a new bias term arising in the asymptotic approximations. The following assumption

collects the restrictions we impose on the first-step estimation procedure.

Assumption 1 (First-Step Estimate)**.** (i) $\max_{1\leq i\leq n}|\hat{\mu}_i-\mu_i| = o_{\mathbb{P}}(1)$, and (ii) $\max_{1\leq i\leq n}|\eta_i| = o_{\mathbb{P}}(1/\sqrt{n})$.

This assumption imposes high-level conditions on the first-step estimate, covering both series-based nonparametric estimation and, more generally, many covariates settings. Assumption 1(i) requires uniform consistency of $\hat{\mu}_i$ for μ_i. In Section SA-2 of the Supplemental Appendix we discuss primitive conditions in different scenarios, covering (i) series-based estimation, (ii) generic covariates with bounded higher moments, (iii) generic covariates with alternative conditions on the tails of their distribution, and (iv) generic covariates formed using many dummy/discrete regressors. The list is not meant to be exhaustive, and primitive conditions for other cases can be found in the vast literatures on nonparametric sieve estimation and high-dimensional models.

Assumption 1(ii) requires the linear approximation error in (6) be small relative to the sample size. The latter assumption is standard in two-step semiparametric settings, where the covariates $\mathbf{z}_i$ include basis expansions able to approximate $\mu_i = \mu(\mathbf{z}_i)$ accurate enough. This assumption also holds easily when covariates are discrete and a fully saturated model is used. However, Assumption 1(ii) cannot be dropped without affecting the interpretation of the final estimand $\boldsymbol{\theta}_0$ because the best linear approximation in the first step will affect (in general) the probability limit of the resulting two-step estimator. In other words, either the researcher assumes that the best linear approximation is approximately exact in large samples, or needs to change the interpretation of the probability limit of the two-step estimator because of the misspecification introduced in the first step. The later approach is common in empirical work, where researchers often employ a "flexible" parametric model, such as linear regression, Probit or Logit, all of which are misspecified in general.

Remark 1 (Extensions)**.** In the Supplemental Appendix, we extend our main results in three directions. First, in Section SA-4.1 we allow for a multidimensional first-step $\boldsymbol{\mu}_i$ entering the second-step estimating equation $\mathbf{m}(\mathbf{w},\cdot,\boldsymbol{\theta})$. Second, in Section SA-4.2 we allow for a partially linear first-step structure as opposed to (6). Both extensions are conceptually straightforward (they require additional notation and tedious algebra), but are nonetheless key to handle the production function example discussed in Section 6.1. Finally, in Section SA-4.3 we discuss an special case of two-step estimators where high-dimensional covariates enter both the first-step (through μ_i) and

the second-step (in an additively separable way). This extension is useful in the context of marginal treatment effect estimation and inference, as we illustrate in Section 6.2. Allowing for the high-dimensional covariates to enter the second-step estimating equation in an unrestricted way makes the problem quite difficult, and therefore we relegate the general case for future work. ⌟

3.2 Distribution Theory

It is not difficult to establish $\hat{\boldsymbol{\theta}} \to_{\mathbb{P}} \boldsymbol{\theta}_0$, even when $k/\sqrt{n} = O(1)$. See the Supplemental Appendix for exact regularity conditions. On the other hand, the $\sqrt{n}$-scaled mean squared error and distributional properties of the estimator $\hat{\boldsymbol{\theta}}$ will change depending on whether k is "small" or "larger" relative to the sample size. To describe heuristically the result, consistency of $\hat{\boldsymbol{\theta}}$ and a second-order Taylor series expansion give:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \tag{8}$$

$$+ \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big) \tag{9}$$

$$+ \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big)^2, \tag{10}$$

where $\boldsymbol{\Sigma}_0 = -(\mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_0\mathbf{M}_0)^{-1}\mathbf{M}_0^{\mathsf{T}}\boldsymbol{\Omega}_0$ with $\mathbf{M}_0 = \mathbb{E}\left[\partial\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$, and

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

Term (8) will be part of the influence function. Using conventional large sample approximations (i.e., k fixed or at most $k/\sqrt{n} \to 0$), term (9) contributes to the variability of $\hat{\boldsymbol{\theta}}$ as a result of estimating the first step, and term (10) will be negligible. Here, however, we show that under the many covariates assumption $k/\sqrt{n} \not\to 0$, both (9) and (10) will deliver nonvanishing bias terms. The main intuition is as follows: as the number of covariates increases relative to the sample size, the error in $\hat{\mu}_i - \mu_i$ increases due to overfitting, and translates into less accurate approximations in terms (9) and (10). This, in turn, affects the finite sample performance of the usual asymptotic approximations, delivering unreliable results in applications. To be specific, the term (9) contributes a leave-in bias arising from using the same observation to estimate μ_i and later the parameter $\boldsymbol{\theta}_0$,

while the term (10) contributes with a bias arising from averaging (non-linear) squared errors in the approximation of μ_i.

The following theorem formalizes our main finding. The proof relies on several preliminary results given in the Supplemental Appendix. Let $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]^\mathsf{T}$ be the first step included covariates and $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-}\mathbf{Z}^\mathsf{T}$ be the projection matrix with elements $\{\pi_{ij} : 1 \leq i, j \leq n\}$.

Theorem 1 (Asymptotic Normality)**.**
Suppose Assumptions 1 and A-1 hold, $\hat{\boldsymbol{\theta}} \to_{\mathbb{P}} \boldsymbol{\theta}_0$ the unique solution of $\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})] = \mathbf{0}$ and an interior point of Θ, $\mathbf{\Omega}_n \to_{\mathbb{P}} \mathbf{\Omega}_0$ positive definite. If $k = O(\sqrt{n})$, then (2) holds with

$$\mathscr{B} = \mathbf{\Sigma}_0 \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbf{B}_i|\mathbf{Z}], \qquad \mathscr{V} = \frac{1}{n}\mathbf{\Sigma}_0 \left(\mathbb{V}[\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{Z}]] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}[\mathbf{\Psi}_i|\mathbf{Z}] \right) \mathbf{\Sigma}_0,$$

where

$$\mathbf{B}_i = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(r_i - \mu_i)\pi_{ii} + \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \sum_{j=1}^{n} (r_j - \mu_j)^2 \pi_{ij}^2,$$

$$\mathbf{\Psi}_i = \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + \left(\sum_{j=1}^{n} \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)|\mathbf{Z}]\pi_{ij} \right) (r_i - \mu_i).$$

Using well known properties of projection matrices, it follows that $\mathscr{B} = O_{\mathbb{P}}(k/n)$ and non-zero in general, and thus the distributional approximation in Theorem 1 will exhibit a first-order asymptotic bias whenever k is "large" relative to the sample size (e.g., $k \propto \sqrt{n}$). In turn, this result implies that conventional inference procedures ignoring this first-order distributional bias will be invalid, leading to over-rejection of the null hypothesis of interest and under-coverage of the associated confidence intervals. Section 7 presents simulation evidence capturing this phenomena.

To understand the implications of the above theorem, we discuss the two terms in $\mathbf{B}_i$. The first term corresponds to the contribution from (9), because a first order approximation gives $\mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \approx \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(\hat{\mu}_i - \mu_i) \approx \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(\sum_j \pi_{ij}(r_j - \mu_j))$. Because $\mathbb{E}[r_j - \mu_j|\mathbf{z}_j] = 0$, this bias is proportional to the sample average of $\mathbb{C}\mathrm{ov}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0), r_i - \mu_i|\mathbf{z}_i]\pi_{ii}$. Hence the bias, due to the linear contribution of $\hat{\mu}_i$, will be zero if there is no residual variation in the sensitivity measure $\dot{\mathbf{m}}$ (i.e., $\mathbb{V}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i] = \mathbf{0}$) or, more generally, the residual variation in the sensitivity measure $\dot{\mathbf{m}}$ is uncorrelated to the first step error term (i.e., $\mathbb{C}\mathrm{ov}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0), r_i - \mu_i|\mathbf{z}_i] = \mathbf{0}$ for all $i = 1, 2, \ldots, n$).

The second term in $\mathbf{B}_i$ captures the quadratic dependence of the estimating equation on the unobserved μ_i, coming from (10). Because of the quadratic nature, this bias represents the accumulated estimation error when $\hat{\mu}_i$ is overfitted. When $i \neq j$, which is the main part of the bias, $\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(r_j - \mu_j)^2 | \mathbf{z}_i, \mathbf{z}_j] = \mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]\mathbb{E}[(r_j - \mu_j)^2|\mathbf{z}_j]$, and hence this portion of the bias will be non-zero unless an estimating equation linear in μ_i is considered or, slightly more generally, $\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i] = \mathbf{0}$. Intuitively, overfitting the first step does not give a quadratic contribution if the estimating equation is not sensitive on average to the second order (at μ_i).

The first bias could be manually removed by employing a leave-one-out estimator of μ_i. However, the second bias cannot be removed this way. Furthermore, the leave-one-out estimator $\hat{\mu}_i^{(i)}$ usually has higher variability compared with $\hat{\mu}_i$, hence the second bias will be amplified, which is confirmed by our simulations.

Chernozhukov, Escanciano, Ichimura, and Newey (2017) introduced the class of locally robust estimators, which are a generalization of doubly robust estimators (Bang and Robins, 2005) and the efficient influence function estimators (Cattaneo, 2010, p. 142). These estimators can offer demonstrable improvements in terms of smoothing/approximation bias rate restrictions and, consequently, they offer robustness to "small" k (underfitting). See also Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2017) and Newey and Robins (2017) for related approaches. This type of estimators are carefully constructed so that (9) is removed, but they do not account for (10). Because the "large" k bias is in part characterized by (10), locally robust estimators cannot (in general) reduce the bias we uncover in this paper. Therefore, our methods complement locally robust estimation by offering robustness to overfitting, that is, situations where the first step estimate includes possibly many covariates.

Consider next the variance and distributional approximation. Theorem 1 shows that the distributional properties of $\hat{\boldsymbol{\theta}}$ are based on a double sum in general, and hence it does not have an "influence function" or asymptotically linear representation. Nevertheless, after proper Studentization, asymptotic normality holds as in (2). The following remark summarizes the special case when the estimator, after bias correction, does have an asymptotic linear representation.

Remark 2 (Asymptotic Linear Representation)**.** Suppose the conditions of Theorem 1 hold and, in addition, $\inf_{\boldsymbol{\Gamma}} \mathbb{E}[|\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i] - \boldsymbol{\Gamma}\mathbf{z}_i|^2] \to 0$ as $k \to \infty$, that is, $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ can be

well approximated by a linear combination of $\mathbf{z}_i$ in mean square. Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \mathscr{B}) = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) | \mathbf{z}_i](r_i - \mu_i) \Big\} + o_{\mathbb{P}}(1),$$

hence $\hat{\boldsymbol{\theta}}$ is asymptotically linear after bias correction even when $k/\sqrt{n} \not\to 0$. However, $\hat{\boldsymbol{\theta}}$ is asymptotically linear if and only if $k/\sqrt{n} \to 0$ in general. See Newey (1994) and Hahn and Ridder (2013) for more discussion on asymptotic linearity and variance calculations. ⌟

In practice one needs to estimate both the asymptotic bias and the variance to conduct valid statistical inference. Plug-in estimators could be constructed to this end, though additional unknown functions would need to be estimated (e.g., conditional expectations of derivatives of the estimating equation). Under appropriate regularity conditions, these estimators would be consistent for the bias and variance terms. As a practically relevant alternative, we show in the upcoming sections that the jackknife can be used to estimate both the bias and variance, and that a carefully crafted resampling method can be used to conduct inference. The key advantage of these results is that they are fully automatic, and therefore can be used for any model considered in practice without having to re-derive and plug-in for the exact expressions each time.

Remark 3 (Delta Method)**.** Our results apply directly to many other estimands via the so-called delta method. Let $\boldsymbol{\varphi}(\cdot)$ be a possibly vector-valued continuously differentiable function of the parameter $\boldsymbol{\theta}_0$ with gradient $\dot{\boldsymbol{\varphi}}(\cdot)$. Then, under the conditions of Theorem 1,

$$\Big(\dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0) \mathscr{V} \dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^{\mathsf{T}} \Big)^{-1/2} \Big(\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\varphi}(\boldsymbol{\theta}_0) - \dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0) \mathscr{B} \Big) \rightsquigarrow \text{Normal}(\mathbf{0}, \mathbf{I}),$$

provided that $\dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)$ is full rank. Hence, the usual delta method can be used for estimation and inference, even the two-step GMM setting with a high-dimensional first-step estimate. ⌟

Plug-in consistent estimation of the appropriate GMM efficient weighting matrix is also possible given our regularity conditions, but we do not give details here to conserve space. An interesting extension of our work would be to consider non-differentiable second-step estimating equations as in, for example, Abadie, Angrist, and Imbens (2002) or Cattaneo (2010) for quantile and related treatment effect models.

4 Jackknife Bias Correction and Variance Estimation

We show that the jackknife is able to estimate consistently the many covariate bias and the asymptotic variance of $\hat{\boldsymbol{\theta}}$, even when $k = O(\sqrt{n})$, and without assuming a valid asymptotic linear representation for $\hat{\boldsymbol{\theta}}$.

The jackknife estimates are constructed by simply deleting one observation at the time and then re-estimating both the first and second steps. To be more specific, let $\hat{\mu}_i^{(\ell)}$ denote the first-step estimate after the ℓth observation is removed from the dataset. Then, the leave-ℓ-out two-step estimator is

$$\hat{\boldsymbol{\theta}}^{(\ell)} = \arg\min_{\boldsymbol{\theta}} \left| \boldsymbol{\Omega}_n^{1/2} \sum_{i=1, i\neq \ell}^{n} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{(\ell)}, \boldsymbol{\theta}) \right|, \qquad \ell = 1, 2, \ldots, n.$$

Finally, the bias and variance estimates are constructed as in (4). This approach is fully data-driven and automatic. The Supplemental Appendix (Section SA-5) gives further details on implementation.

In addition to being fully automatic, another appealing feature of the jackknife in our case is that it is possible to exploit the specific structure of the problem to reduce computation burden. Specifically, because we consider a linear regression fit for the first step, the leave-ℓ-out estimate $\hat{\mu}_i^{(\ell)}$ can easily be obtained by

$$\hat{\mu}_i^{(\ell)} = \hat{\mu}_i + \frac{\hat{\mu}_\ell - r_\ell}{1 - \pi_{\ell\ell}} \cdot \pi_{i\ell}, \qquad 1 \leq i \leq n,$$

where recall that $\pi_{i\ell}$ is the $(i, \ell)^{\text{th}}$ element of the projection matrix for the first step $\boldsymbol{\Pi} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^- \mathbf{Z}^\top$. Since recomputing the first-step estimate can be time-consuming when k is large, the above greatly simplifies the algorithm and reduces computing time.

To show the validity of the jackknife, we impose two additional mild assumptions on the possibly large dimensional covariates $\mathbf{z}_i$, captured through the projection matrix of the first-step estimate.

Theorem 2 (Jackknife-Based Valid Inference)**.**
Suppose the conditions of Theorem 1 hold. If, in addition, (i) $\sum_{1\leq i\leq n} \pi_{ii}^2 = o_{\mathbb{P}}(k)$ and (ii) $\max_{1\leq i\leq n} 1/(1 - \pi_{ii}) = O_{\mathbb{P}}(1)$, then (3) holds.

The two conditions together correspond to "design balance", which states that asymptotically

the projection matrix is not "concentrated" on a few observations. They are slightly weaker than $\max_{1\leq i\leq n}\pi_{ii} = o_{\mathbb{P}}(1)$, which is commonly assumed in the literature on high-dimensional statistics. In Section SA-2.4 of the Supplemental Appendix we give a concrete example using sparse dummy covariates where $\max_{1\leq i\leq n}\pi_{ii} \neq o_{\mathbb{P}}(1)$, but the conditions (i) and (ii) are satisfied. For more discussion on design balance in linear least squares models see, e.g., Chatterjee and Hadi (1988).

Remark 4 (Delta Method)**.** Consider the setup of Remark 3, where the goal is to conduct estimation and inference for a (smooth) function of $\boldsymbol{\theta}_0$. In this case, the estimator is $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}})$. There are at least three ways to conduct bias correction: (i) plug-in method leading to $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}} - \hat{\mathscr{B}})$, (ii) linearization-based method leading to $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}}) - \dot{\boldsymbol{\varphi}}(\hat{\boldsymbol{\theta}})\hat{\mathscr{B}}$, and (iii) direct jackknife of $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}})$. The three methods are asymptotically equivalent, and can be easily implemented in practice. The same argument applies to the variance estimator when $\boldsymbol{\varphi}(\boldsymbol{\theta}_0)$ is the target parameter. ⌟

By showing the validity of the jackknife, one can construct confidence intervals and conduct hypothesis tests using the jackknife bias and variance estimators, and the normal approximation. In particular, bias correction will not affect the variance of the asymptotic distribution. On the other hand, any bias correction technique is likely to introduce additional variability, which can be nontrivial in finite samples. This is indeed confirmed by our simulation studies. In the next section, we introduce a carefully crafted fully automatic bootstrap method that can be applied to the bias-corrected Studentized statistic to obtain better finite sample distributional approximations.

5 Bootstrap Inference after Bias Correction

In this section we develop a fast, automatic and specifically tailored bootstrap-based approach to conducting post-bias-correction inference in our setting. The method combines the wild bootstrap (first-step estimation) and the multiplier bootstrap (second-step estimation) to give an easy-to-implement valid distributional approximation to the finite sample distribution of the jackknife-based bias-corrected Studentized statistic in (3). See Mammen (1993) for a related result in the context of a high-dimensional one-step linear regression model without any bias-correction, and Kline and Santos (2012) for some recent higher-order results in the context of parametric low-dimensional linear regression models.

Let $\{\omega_i^\star : 1 \le i \le n\}$ be i.i.d. bootstrap weights with $\mathbb{E}[\omega_i^\star] = 1$, $\mathbb{V}[\omega_i^\star] = 1$ and $\mathbb{E}[(\omega_i^\star - 1)^3] = 0$. First, we describe the bootstrap construction of $\hat{\boldsymbol{\theta}}^\star$. We employ the wild bootstrap to obtain $\hat{\mu}_i^\star$, mimicking the first-step estimate (7): we regress $r_i^\star$ on $\mathbf{z}_i$, where $r_i^\star = \hat{\mu}_i + (\omega_i^\star - 1)(r_i - \hat{\mu}_i)$. Then, we employ the multiplier bootstrap to obtain $\hat{\boldsymbol{\theta}}^\star$, mimicking the second-step estimate (1):

$$\hat{\boldsymbol{\theta}}^\star = \arg\min_{\boldsymbol{\theta}} \left| \boldsymbol{\Omega}_n^{1/2} \sum_{i=1}^n \omega_i^\star \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^\star, \boldsymbol{\theta}) \right|. \tag{11}$$

Second, we describe the bootstrap construction of $\hat{\mathscr{B}}$ and $\hat{\mathscr{V}}$; that is, the implementation of the jackknife bias and variance estimators under the bootstrap. Because we employ a multiplier bootstrap, the jackknife estimates need to be adjusted to account for the effective number of observations under the bootstrap law. Thus, we have:

$$\hat{\mathscr{B}}^\star = (n-1)(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}^\star), \qquad \hat{\mathscr{V}}^\star = \frac{n-1}{n} \sum_{\ell=1}^n \omega_\ell^\star (\hat{\boldsymbol{\theta}}^{\star,(\ell)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)})(\hat{\boldsymbol{\theta}}^{\star,(\ell)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)})^\mathsf{T},$$

where

$$\hat{\boldsymbol{\theta}}^{\star,(\cdot)} = \frac{1}{n} \sum_{\ell=1}^n \omega_\ell^\star \hat{\boldsymbol{\theta}}^{\star,(\ell)},$$

$$\hat{\boldsymbol{\theta}}^{\star,(\ell)} = \arg\min_{\boldsymbol{\theta}} \left| \boldsymbol{\Omega}_n^{1/2} \Big\{ \sum_{i=1}^n \big[\omega_i^\star - \mathbb{1}_{(i=\ell)}\big] \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{\star,(\ell)}, \boldsymbol{\theta}) \Big\} \right|, \qquad \ell = 1, 2, \ldots, n.$$

Here $\hat{\mu}_i^{\star,(\ell)}$ is obtained by regressing $r_i^\star$ on $\mathbf{z}_i$, without using the ℓth observation. Equivalently, the jackknife deletes the ℓth observation in the first step wild bootstrap, and reduces the ℓth weight $\omega_\ell^\star$ by 1 in the second step multiplier bootstrap.

Our resampling approach employs the wild bootstrap to form $\hat{\mu}_i^\star$, which is very easy and fast to implement and does not require recomputing the possibly high-dimensional projection matrix $\boldsymbol{\Pi}$, and then uses the same bootstrap weights to construct $\hat{\boldsymbol{\theta}}^\star$ via a multiplier resampling approach. It is possible to use the multiplier bootstrap for both estimation steps, which would give a more unified treatment, but such an approach is harder to implement and does not utilizes efficiently (from a computational point of view) the specific structure of the first-step estimate. To be more specific, employing the multiplier bootstrap in the first-step estimation leads to $\hat{\mu}_i^\star = \mathbf{z}_i^\mathsf{T}(\mathbf{Z}^\mathsf{T}\mathbf{W}^\star\mathbf{Z})^- \mathbf{Z}^\mathsf{T}\mathbf{W}^\star\mathbf{R}$, where $\mathbf{R} = [r_1, r_2, \ldots, r_n]^\mathsf{T}$ and $\mathbf{W}^\star$ is a diagonal matrix with diagonal elements $\{\omega_i^\star\}_{1 \le i \le n}$, which

requires recomputing the projection matrix for each bootstrap replication. In contrast, our bootstrap approach leads to $\hat{\mu}_i^\star = \mathbf{z}_i^\mathsf{T}(\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-}\mathbf{Z}^\mathsf{T}\mathbf{R}^\star$, where $\mathbf{R}^\star = [r_1^\star, r_2^\star, \ldots, r_n^\star]^\mathsf{T}$. As discussed before, this important practical simplification also occurs because we are employing a linear regression fit in the first step. Employing the standard nonparametric bootstrap may also be possible, but additional (stronger) regularity conditions would be required. Last but not least, we note that combining the jackknife with the multiplier bootstrap naïvely (that is, deleting the ℓth observation with its weight $\omega_\ell^\star$ altogether in the second step) does not deliver a consistent variance estimate; see the Supplemental Appendix for details.

The following theorem summarizes our main result for inference. Only two additional mild, high-level conditions on the bootstrap analogue first-step and second-step estimators are imposed.

Theorem 3 (Bootstrap Validity)**.**
Suppose the conditions of Theorems 1 and 2 hold. If, in addition, $\max_{1\le i\le n}|\hat{\mu}_i^\star - \hat{\mu}_i| = o_{\mathbb{P}}(1)$ and $|\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}}| = o_{\mathbb{P}}(1)$, then (5) holds.

It is common to assume the bootstrap weights $\omega_i^\star$ to have mean 1 and variance 1. For the jackknife bias and variance estimator to be consistent under the bootstrap distribution, we also need that the third central moment of $\omega_i^\star$ is zero. Examples include $\omega_i^\star = 1 + e_i^\star$ with $e_i^\star$ following the Rademacher distribution or the six-point distribution proposed in Webb (2014).

For inference, consider for example the one dimensional case: $\dim(\boldsymbol{\theta}_0) = 1$. The bootstrap percentile-t bias-corrected (equal tail) confidence interval for θ_0 is

$$\left[\ \hat{\theta} - \hat{\mathscr{B}} - \hat{q}_{1-\alpha/2}\cdot\sqrt{\hat{\mathscr{V}}}\quad,\quad \hat{\theta} - \hat{\mathscr{B}} - \hat{q}_{\alpha/2}\cdot\sqrt{\hat{\mathscr{V}}}\ \right],$$

where $\hat{q}_\alpha = \inf\{t\in\mathbb{R} : \hat{F}(t)\ge\alpha\}$ is the empirical α^{th} quantile of $\{\mathscr{T}_b^\star : 1\le b\le B\}$, with $\hat{F}(t) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}[\mathscr{T}_b^\star\le t]$ and $\mathscr{T}_b^\star$ denoting the bootstrap statistic in (5) in b^{th} simulation.

6 Examples

We now apply our results to two leading examples: production function estimation, which is an IO application with multidimensional partially linear first step estimation, and marginal treatment effect, which relates to IV methods with heterogeneous treatment effects and many covariates. The

Supplemental Appendix analyzes several other examples of interest in applied microeconomics. For each example, besides the general form of the bias in Theorem 1, we show that it is possible to further characterize the nature and source of the many covariates bias by utilizing corresponding identification assumptions in each of the examples.

6.1 Production Function

As a first substantive application of our results to empirically relevant problems in applied microeconomics, we consider production function estimation. For a review of this topic, including an in-depth discussion of its applicability to industrial organization and other fields in Economics, see Ackerberg, Benkard, Berry, and Pakes (2007). For concreteness, here we focus on the setting introduced by Olley and Pakes (1996), and propose new estimation and inference methods for production functions allowing for possibly many covariates in the first step estimation. To apply our methods to this problem, two extensions mentioned previously (multidimensional and partially linear first step estimation) are needed, and therefore the results below are notationally more involved.

We use i to index firms (i.e., observations) and t for time. The production function takes the form:

$$Y_{i,t} = \beta_L L_{i,t} + \beta_K K_{i,t} + \beta_A A_{i,t} + W_{i,t} + U_{i,t}, \tag{12}$$

where Y, L, K and A represent (log) output, labor input, capital input and aging effect, respectively. W is the firm-specific productivity factor, and is a (generalized) fixed effect. The error term U is either measurement error or shock that is unpredictable with time-t information, hence has zero conditional mean (given right-hand-side variables). Since that the productivity factor W is unobserved, (12) cannot be used directly to estimate the production function.

Now we discuss briefly the decision process of a firm in each period. First, the firm compares continuation value to salvage (liquidation) value, and decides whether or not to exit the market. Upon deciding to stay in business, the firm chooses simultaneously the labor input $L_{i,t}$ and investment $I_{i,t}$, given its private knowledge about productivity $W_{i,t}$. Finally, capital stock follows the classical law of motion.

The first crucial assumption for identification is that there is a one-to-one relationship between

the firm-level decision variable $I_{i,t}$ and the unobserved state variable $W_{i,t}$, which allows inverting the investment decision and write $W_{i,t} = h_t(I_{i,t}, K_{i,t}, A_{i,t})$, with h_t unknown and possibly time-dependent. Then, (12) can be written as

$$Y_{i,t} = \beta_L L_{i,t} + \phi_{i,t} + U_{i,t}, \qquad \phi_{i,t} = \beta_K K_{i,t} + \beta_A A_{i,t} + h_t(I_{i,t}, K_{i,t}, A_{i,t}). \tag{13}$$

The above equation can be used to estimate the labor share β_L flexibly using a partially linear regression approach. The capital share β_K and the effect of aging β_A, however, are not identified.

To identify β_K and β_A, we need to use information embedded in firm's exit decision. Olley and Pakes (1996) showed that firm's exit decision can be summarized as $\chi_{i,t} = \mathbb{1}[W_{i,t} \geq \underline{W}_t(A_{i,t}, K_{i,t})]$, where $\chi_{i,t} = 1$ represents the firm staying in business, and $\underline{W}_t$ is the threshold function. Then, we decompose $W_{i,t+1}$ into

$$W_{i,t+1} = \mathbb{E}[W_{i,t+1}|W_{i,t}, \chi_{i,t+1} = 1] + V_{i,t+1}.$$

Conditioning on survival at time $t+1$ (i.e. $\chi_{i,t+1} = 1$) is the same as conditioning on $\underline{W}_{i,t}$, and hence the conditional expectation in the above display is an unknown function of $W_{i,t}$ and $\underline{W}_{i,t}$.

The second crucial assumption for identification is that the survival probability, defined as

$$P_{i,t} = P_t(I_{i,t}, K_{i,t}, A_{i,t}) = \mathbb{E}[\chi_{i,t+1}|I_{i,t}, K_{i,t}, A_{i,t}], \tag{14}$$

is a valid proxy for $\underline{W}_{i,t}$. Therefore, with the time index progressed by one period, we rewrite (12) as

$$\begin{aligned} Y_{i,t+1} - \beta_L L_{i,t+1} &= \beta_K K_{i,t+1} + \beta_A A_{i,t+1} + g(P_{i,t},\ W_{i,t}) \qquad\qquad + V_{i,t+1} + U_{i,t+1} \\ &= \beta_K K_{i,t+1} + \beta_A A_{i,t+1} + g(P_{i,t},\ \phi_{i,t} - \beta_K K_{i,t} - \beta_A A_{i,t}) + V_{i,t+1} + U_{i,t+1}. \end{aligned} \tag{15}$$

Here we make an important remark on the two error terms and why labor input has been moved to the left-hand-side, which also sheds light on the estimation strategy. $U_{i,t+1}$ is either a measurement error or the conditional expectation error of $Y_{i,t+1}$ on contemporaneous variables, hence is orthogonal to time-$t+1$ information. On the other hand, $V_{i,t+1}$ is the conditional expectation error of $\chi_{i,t+1}$ on time-t variables, hence is only orthogonal to time-t information. It is uncorrelated with

$K_{i,t+1}$ and $A_{i,t+1}$ since they are predetermined, but in general *correlated* with $L_{i,t+1}$. This is the endogeneity problem underlying (15), and shows why it cannot be used for estimation without β_L being estimated in a first step.

Now we describe the estimation strategy. For simplicity we make two assumptions: (i) there are only two periods $t = 1, 2$, and (ii) the function g is known up to a finite dimensional parameter $\boldsymbol{\lambda}_0$. First, we rely on (13) to estimate β_L and $\phi_{i,1}$ with a partially linear regression, which gives $\hat{\beta}_L$ and $\hat{\phi}_{i,1}$. Second, we use (14) to obtain the estimated probability of staying, $\hat{P}_{i,1}$. These are the two first-step estimates in this application. Finally, given the preliminary estimates, β_K, β_L and the nuisance parameter $\boldsymbol{\lambda}_0$ are jointly estimated in the second step. The entire two-step estimation approach is summarized as follows:

$$\begin{bmatrix} \hat{\beta}_K \\ \hat{\beta}_A \\ \hat{\boldsymbol{\lambda}} \end{bmatrix} = \underset{\beta_K, \beta_A, \boldsymbol{\lambda}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[Y_{i,2} - \hat{\beta}_L L_{i,2} - \beta_K K_{i,2} - \beta_A A_{i,2} - g(\hat{P}_{i,1}, \hat{\phi}_{i,1} - \beta_K K_{i,1} - \beta_A A_{i,1}, \boldsymbol{\lambda}) \right]^2,$$

$$\hat{\phi}_{i,1} = \mathbf{Z}_{i,1}^{\mathsf{T}} \hat{\boldsymbol{\gamma}}_1, \qquad [\hat{\beta}_L, \hat{\boldsymbol{\gamma}}_1^{\mathsf{T}}]^{\mathsf{T}} = \underset{\beta, \boldsymbol{\gamma}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_{i,1} - \beta L_{i,1} - \mathbf{Z}_{i,1}^{\mathsf{T}} \boldsymbol{\gamma} \right)^2,$$

$$\hat{P}_{i,1} = \mathbf{Z}_{i,1}^{\mathsf{T}} \hat{\boldsymbol{\gamma}}_2, \qquad \hat{\boldsymbol{\gamma}}_2 = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\chi_{i,2} - \mathbf{Z}_{i,1}^{\mathsf{T}} \boldsymbol{\gamma} \right)^2,$$

with $\mathbf{Z}_{i,1}$ being series expansion based on the variables $(I_{i,1}, K_{i,1}, A_{i,1})$, in addition to perhaps other variables.

The estimation problem does not fit into our basic framework for three reasons. First, we have two estimating equations in the first step, one for $(\beta_L, \phi_{i,1})$ and the other for $P_{i,1}$. Second, the model features a parameter (β_L) estimated in the first step and then plugged into the second step estimating equation. Third, $\hat{\phi}_{i,1}$ is no longer a conditional expectation projection, but is instead obtained from a partially linear regression. As mentioned above, Section SA-4 in the Supplemental Appendix discusses extensions of our framework that enable us to handle this application in full generality.

Applying Theorem 1, properly extended using the results in Section SA-4 of the Supplemental Appendix, we have the following results on bias and variance for the estimator $[\hat{\beta}_K, \hat{\beta}_A, \hat{\boldsymbol{\lambda}}^{\mathsf{T}}]^{\mathsf{T}}$.

Corollary 1 (Asymptotic Normality: Production Function)**.**

Assume the assumptions of Theorem 1 and the example-specific additional regularity conditions summarized in the Supplemental Appendix hold. Then,

$$\mathbf{B}_i = \Big(\mathbf{b}_{1,i} + \mathbf{b}_{2,i}\Big)\pi_{ii} + \sum_{j=1}^{n}\Big(\mathbf{b}_{3,ij} + \mathbf{b}_{4,ij} + \mathbf{b}_{5,ij}\Big)\pi_{ij}^2$$

$$\mathbf{\Psi}_i = \mathbf{\Psi}_{1,i} + \mathbf{\Psi}_{2,i} + \mathbf{\Psi}_{3,i},$$

where

$$\mathbf{b}_{1,i} = \begin{bmatrix} K_{i,1}g_{22,i} \\ A_{i,1}g_{22,i} \\ -\mathbf{g}_{23,i} \end{bmatrix} U_{i,1}V_{i,2}, \qquad \mathbf{b}_{2,i} = \begin{bmatrix} K_{i,1}g_{12,i} \\ A_{i,1}g_{12,i} \\ -\mathbf{g}_{13,i} \end{bmatrix} (\chi_{i,2} - P_{i,1})V_{i,2}$$

$$\mathbf{b}_{3,ij} = \left(- \begin{bmatrix} K_{i,1}g_{22,i} \\ A_{i,1}g_{22,i} \\ -\mathbf{g}_{23,i} \end{bmatrix} g_{2,i} - \frac{1}{2} \begin{bmatrix} K_{i,1}g_{2,i} - K_{i,2} \\ A_{i,1}g_{2,i} - A_{i,2} \\ -\mathbf{g}_{3,i} \end{bmatrix} g_{22,i} \right) U_{i,1}^2,$$

$$\mathbf{b}_{4,ij} = \left(- \begin{bmatrix} K_{i,1}g_{12,i} \\ A_{i,1}g_{12,i} \\ -\mathbf{g}_{13,i} \end{bmatrix} g_{1,i} - \frac{1}{2} \begin{bmatrix} K_{i,1}g_{2,i} - K_{i,2} \\ A_{i,1}g_{2,i} - A_{i,2} \\ -\mathbf{g}_{3,i} \end{bmatrix} g_{11,i} \right) (\chi_{j,2} - P_{j,1})^2,$$

$$\mathbf{b}_{5,ij} = \left(\begin{bmatrix} K_{i,1}g_{22,i} \\ A_{i,1}g_{22,i} \\ -\mathbf{g}_{23,i} \end{bmatrix} g_{1,i} + \begin{bmatrix} K_{i,1}g_{12,i} \\ A_{i,1}g_{12,i} \\ -\mathbf{g}_{13,i} \end{bmatrix} g_{2,i} + \begin{bmatrix} K_{i,1}g_{2,i} - K_{i,2} \\ A_{i,1}g_{2,i} - A_{i,2} \\ -\mathbf{g}_{3,i} \end{bmatrix} g_{12,i} \right) (\chi_{j,2} - P_{j,1})U_{j,1}$$

$$\mathbf{\Psi}_{1,i} = \begin{bmatrix} K_{i,1}g_{2,i} - K_{i,2} \\ A_{i,1}g_{2,i} - A_{i,2} \\ -\mathbf{g}_{3,i} \end{bmatrix} \Big(V_{i,2} + U_{i,2}\Big)$$

$$\mathbf{\Psi}_{2,i} = - \begin{bmatrix} K_{i,1}g_{2,i} - K_{i,2} \\ A_{i,1}g_{2,i} - A_{i,2} \\ -\mathbf{g}_{3,i} \end{bmatrix} g_{2,i}U_{i,1} - \begin{bmatrix} K_{i,1}g_{2,i} - K_{i,2} \\ A_{i,1}g_{2,i} - A_{i,2} \\ -\mathbf{g}_{3,i} \end{bmatrix} g_{1,i}\Big(\chi_{i,2} - P_{i,1}\Big)$$

$$\mathbf{\Psi}_{3,i} = \frac{1}{\mathbb{E}[\mathbb{V}[L_{i,1}|\mathbf{Z}_{i,1}]]}\mathbf{\Xi}_0\Big(L_{i,1} - \mathbb{E}[L_{i,1}\Big|\mathbf{Z}_{i,1}]\Big)U_{i,1}.$$

We use the abbreviation $g_i = g(P_{i,1}, W_{i,1}, \boldsymbol{\lambda}_0)$, and further subscripts 1, 2 and 3 of g_i are used to

denote its partial derivatives with respect to the first, second and third argument, respectively. Exact formulas of $\mathbf{\Sigma}_0$ and $\mathbf{\Xi}_0$ are available in the Supplemental Appendix (Section SA-4.2 and SA-5.7).

Some bias terms can be made to zero with additional assumptions. First consider the scenario that $U_{i,t}$ is purely measurement error. Then it should be independent of other error terms, which implies $\mathbf{b}_{1,i}$ and $\mathbf{b}_{5,ij}$ are zero after taking conditional expectations. Sometimes it is assumed that all firms survive from time-1 to time-2 (i.e., there is no sample attrition), or the analyst focuses on a subsample (in which case the parameters have to be reinterpreted). Then $\chi_{i,2} = P_{i,1} = 1$, hence $\mathbf{b}_{2,i}$ and $\mathbf{b}_{4,ij}$ are zero. The variance term $\mathbf{\Psi}_{i,2}$ also simplifies.

6.2 Marginal Treatment Effect

Originally proposed by Björklund and Moffitt (1987), and later developed and popularized by Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006), the marginal treatment effect (MTE) is an important parameter of interest in program evaluation and causal inference. Not only it can be viewed as a limiting version of the local average treatment effect (LATE) of Imbens and Angrist (1994) for continuous instrumental variables (c.f. Angrist, Graddy, and Imbens, 2000), but also it can be used to unify and interpret many other treatment effects parameters such as the average treatment effect or the treatment effect on the treated. Another appealing feature of the MTE is that it provides a description of treatment effect heterogeneity.

To describe the MTE, we adopt a potential outcomes framework under random sampling. Suppose $(Y_i, T_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, 2, \ldots, n$, is i.i.d., where Y_i is the outcome of interest, T_i is a treatment status indicator, $\mathbf{X}_i \in \mathbb{R}^{d_x}$ is a d_x-variate vector of observable characteristics, and $\mathbf{Z}_i \in \mathbb{R}^k$ is k-variate vector of "instruments" (which may include $\mathbf{X}_i$ and transformations thereof). The observed data is generated according to the following switching regression model, also known as potential outcomes or the Roy model,

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0), \qquad Y_i(1) = g_1(\mathbf{X}_i) + U_{1i}, \qquad Y_i(0) = g_0(\mathbf{X}_i) + U_{0i}, \tag{16}$$

$$T_i = \mathbb{1}[P_i \geq V_i], \qquad P_i = P(\mathbf{Z}_i) = \mathbb{E}[T_i|\mathbf{Z}_i], \qquad V_i|\mathbf{X}_i \sim \text{Uniform}[0, 1], \tag{17}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes when an individual receives the treatment or

not, (U_{1i}, U_{0i}, V_i) are unobserved error terms, and P_i is the propensity score or probability of selection. The selection equation (17) is taken essentially without loss of generality to be of the single threshold-crossing form (see Vytlacil, 2002, for more discussion), thought this representation may affect the interpretation of the unobserved heterogeneity.

The (conditional on $\mathbf{X}_i$) MTE at level a is defined as

$$\tau_{\texttt{MTE}}(a|\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0)|V_i = a, \mathbf{X}_i = \mathbf{x}].$$

The MTE will be constant in a if either (i) the individual treatment effect $Y_i(1) - Y_i(0)$ is constant, or (ii) there is no selection on unobservables, that is, the error terms of the outcome equation (16) are unrelated to that of the selection equation (17). The parameter $\tau_{\texttt{MTE}}(a|\mathbf{x})$ is understood as the treatment effect for an individual who is at the margin. Equivalently, it is the treatment effect for the subpopulation where an infinitesimal increase in the propensity score leads to a change in participation status. Note that for a close to 1, the MTE measures the treatment effect in a subpopulation that is very *unlikely* to be treated. Other treatment and policy effects can be recovered using the MTE.

Two assumptions are made to facilitate identification. First, the collection of instruments $\mathbf{Z}_i$ is nondegenerate and independent of the error terms (U_{1i}, U_{0i}, V_i) conditional on the covariates $\mathbf{X}_i$. Second, the probability $\mathbb{P}[T_i = 1|\mathbf{X}_i]$ is bounded away from zero and one. It can then be shown that, for any limit point a in the support of the propensity score, $\tau_{\texttt{MTE}}(a|\mathbf{x})$ is

$$\tau_{\texttt{MTE}}(a|\mathbf{x}) = \frac{\partial}{\partial a}\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}].$$

This representation shows that the MTE is identifiable, and could in principle be estimated by standard nonparametric techniques (once P_i is estimated). In practice, however, nonparametric methods for estimating $\tau_{\texttt{MTE}}(a|\mathbf{x})$ and functionals thereof are often avoided because of the curse of dimensionality, the negative impact of smoothing and tuning parameters, and efficiency considerations. A flexible parametric functional form can be used instead: $\mathbb{E}[Y_i|P_i, \mathbf{X}_i] = e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)$, where $e(\cdot)$ is a known function up to some finite dimensional parameter $\boldsymbol{\theta}_0$.

Therefore, the MTE estimator is often constructed as follows:

$$\hat{\tau}_{\texttt{MTE}}(a|\mathbf{x}) = \frac{\partial}{\partial a} e(\mathbf{x}, a, \hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(Y_i - e(\mathbf{X}_i, \hat{P}_i, \boldsymbol{\theta}) \Big)^2,$$

$$\hat{P}_i = \mathbf{Z}_i^{\mathsf{T}} \hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(T_i - \mathbf{Z}_i^{\mathsf{T}} \boldsymbol{\beta} \Big)^2,$$

Identification and estimation of the MTE, as well as other policy-relevant parameters based on it, require exogenous variation in the treatment equation (17) induced by instrumental variables. In practice, researchers induce this variation by (i) employing many instruments, possibly generating them using power expansions and interactions, and (ii) including interactions with the "raw" or expanded instruments. Employing a flexible, high-dimensional specification for the probability of selection is also useful to mitigate misspecification errors. These observations have led researchers to employ many covariates/instruments in the probability of selection, that is, have a "large" k relative to the sample size. In this paper, we show that flexibly modeling the probability of selection can lead to a first-order bias in the estimation of the MTE and related policy-relevant estimands, even when the outcome equation is modeled parametrically and low-dimensional. Furthermore, we provide automatic bias-correction and inference procedures based on resampling methods.

The following result characterizes the asymptotic properties of the estimated MTE.

Corollary 2 (Asymptotic Normality: MTE)**.**
Suppose the assumptions of Theorem 1 and the example-specific additional regularity conditions summarized in the Supplemental Appendix hold. Then, for $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned}
\mathbf{B}_i &= \frac{\partial^2 e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial P_i \partial \boldsymbol{\theta}} \Big[(1 - P_i) \cdot \mathbb{E}[T_i Y_i(1) | \mathbf{Z}_i] - P_i \cdot \mathbb{E}[(1 - T_i) Y_i(0) | \mathbf{Z}_i] \Big] \pi_{ii} \\
&\quad + \frac{1}{2} \sum_{j=1}^{n} \left[\frac{\partial^2 e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial P_i \partial \boldsymbol{\theta}} \tau_{\texttt{MTE}}(P_i | \mathbf{X}_i) + \frac{1}{2} \frac{\partial e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \tau_{\texttt{MTE}}(P_i | \mathbf{X}_i)}{\partial P_i} \right] P_j (1 - P_j) \pi_{ij}^2, \\
\boldsymbol{\Psi}_i &= \frac{\partial e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \Big) - \left(\sum_{j=1}^{n} \frac{\partial e(\mathbf{X}_j, P_j, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \tau_{\texttt{MTE}}(P_j | \mathbf{X}_j) \pi_{ij} \right) (T_i - P_i),
\end{aligned}$$

and $\boldsymbol{\Sigma}_0$ is given in the Supplemental Appendix (Section SA-5.4).

The above result gives a precise characterization of the asymptotic possibly first-order bias and variance of $\hat{\boldsymbol{\theta}}$ via the results in Theorem 1. To obtain the corresponding result for the estimated

MTE, $\hat{\tau}_{\texttt{MTE}}(a|\mathbf{x})$, the delta method is employed and an extra multiplicative factor $\partial^2 e(\mathbf{x}, a, \boldsymbol{\theta}_0)/\partial a \partial \boldsymbol{\theta}^\mathsf{T}$ shows up. As a result, both the bias and variance for the estimated MTE will depend on the evaluation point $(\mathbf{x}|a)$. We give the details in the Supplemental Appendix Section SA-5.4).

To understand the implications of the above corollary, we consider the bias terms. Note that the factor associated with π_{ii} essentially captures treatment effect heterogeneity (in the outcome equation) and self-selection. To make it zero, one needs to assume there is no heterogeneous treatment effect and the agents do not act on idiosyncratic characteristics that are unobservable to the analyst. For the second bias term associated with π_{ij}^2, note that it involves both the level of the MTE and its curvature. Hence the second bias is related not only to treatment effect heterogeneity captured through the shape of the MTE, but also to the magnitude of the treatment effect. Thus, aside from the off chance of these terms canceling each other, the many instruments bias will be zero only when there is neither heterogeneity nor self-selection, and the treatment effect is zero. Since these conditions are unlikely to hold in empirical work, even in randomized controlled trials, we expect the many instruments bias to have a direct implication in most practical cases. Therefore, conventional estimation and inference methods that do not account for the many instruments bias will be invalid, even in large samples, when many instruments are included in the estimation.

The results above take the conditional expectation function $e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)$ as low-dimensional, but in practice researchers may want to include many covariates also in the second estimation step. In Section SA-4.3 of the Supplemental Appendix, we study a generalization of Theorem 1 for the special case when $\mathbb{E}[Y_i|P_i, \mathbf{X}_i, \mathbf{W}_i] = e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) + \mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}_0$, where $\mathbf{W}_i$ contains additional conditioning variables (possibly including $\mathbf{X}_i$) and the nuisance parameter $\boldsymbol{\gamma}_0$ is potentially high-dimensional. If nonlinear least-squares is used to estimate the second-step, as above, we find that additional terms now contribute to the many covariates bias due to the possibly high-dimensional estimation of $\boldsymbol{\gamma}_0$ in the second-step, but the same general results reported in this paper continue to hold. Specifically, the many covariates bias remains of order $k/\sqrt{n}$ and needs to be accounted for in order to conduct valid inference whenever $k/\sqrt{n}$ is not "small".

7 Numerical Evidence

We provide numerical evidence for the methods developed in this paper. First, we offer a Monte Carlo experiment constructed in the context of MTE estimation (Section 6.2), which highlights the role of the many-covariates bias and showcases the role of jackknife bias correction and bootstrap approximation for estimation and inference. Second, also in the context of MTE estimation and inference, we offer an empirical illustration following the work of Carneiro, Heckman, and Vytlacil (2011). Section SA-8 of the Supplemental Appendix contains more results and further details omitted here to conserve space.

7.1 Simulation Study

We retain the notation and assumptions imposed in Section 6.2, and set the potential outcomes to $Y_i(0) = U_{0i}$ and $Y_i(1) = 0.5 + U_{1i}$. We assume there are many potential instruments $\mathbf{Z}_i = [1, Z_{1,i}, Z_{2,i}, \ldots, Z_{199,i}]$, with $Z_{\ell,i} \sim \text{Uniform}[0,1]$ independent across $\ell = 1, 2, \ldots, 199$. The selection equation is assumed to take a very parsimonious form: $T_i = \mathbb{1}\big[0.1 + Z_{1,i} + Z_{2,i} + Z_{3,i} + Z_{4,i} \geq V_i\big]$. In this case Assumption 1 holds automatically without misspecification error, but in the Supplemental Appendix we explore other specifications of the propensity score where approximation errors are present. 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 $U_{1i}|\mathbf{Z}_i, V_i \sim \text{Uniform}[-0.5, 1.5 - 2V_i]$. Because additional covariates $\mathbf{X}_i$ do not feature in this data generating process, the treatment effect heterogeneity and self-selection are captured by the correlation between U_{1i} and V_i.

It follows that $\mathbb{E}[Y_i|P_i = a] = a - \frac{a^2}{2}$, and the MTE is $\tau_{\texttt{MTE}}(a) = 1 - a$. Given a random sample index by $i = 1, 2, \ldots, n$, the second-step regression model is set to $\mathbb{E}[Y_i|P_i] = \theta_1 + \theta_2 \cdot P_i + \theta_3 \cdot P_i^2$ and therefore the estimated MTE is $\hat{\tau}_{\texttt{MTE}}(a) = \hat{\theta}_2 + 2a \cdot \hat{\theta}_3$ with $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)'$ denoting the least-squares estimators of $(\theta_1, \theta_2, \theta_3)'$. We consider the quantity $\sqrt{n}\,(\hat{\tau}_{\texttt{MTE}}(a) - \tau_{\texttt{MTE}}(a))$ at $a = 0.5$, with and without bias correction, for two sample sizes $n = 1,000$ and $n = 2,000$, and across $2,000$ simulation repetitions. To estimate the propensity score, we regress T_i on a constant term and $\{Z_{\ell,i}\}$ for $1 \leq \ell \leq k-1$, 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.

For inference, we consider two approaches. In the conventional approach, the many instruments

bias is ignored, and hypothesis testing is based on normal approximation to the t-statistic, where the standard error comes from the simulated sampling variability of the estimator (i.e. the oracle standard error, which is infeasible). That is, this benchmark approach considers the infeasible statistic $(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})/\sqrt{\mathbb{V}[\hat{\tau}_{\mathtt{MTE}}]}$, with $\mathbb{V}[\hat{\tau}_{\mathtt{MTE}}]$ denoting the simulation variance of $\hat{\tau}_{\mathtt{MTE}}$, and employs standard normal quantiles. The other approach, which follows the results in this paper, utilizes both the jackknife and the bootstrap: the feasible statistic $(\hat{\tau}_{\mathtt{MTE}} - \hat{\mathscr{B}} - \tau_{\mathtt{MTE}})/\sqrt{\hat{\mathscr{V}}}$ is constructed as in Section 4 and inference is conducted using the bootstrap approximation as in Section 5.

The results are collected in Table 1. The bias is small with small k, as the most parsimonious model is correctly specified. With more instruments added to the propensity score estimation, the many instruments bias quickly emerges, and without bias correction, it leads to severe empirical undercoverage (conventional 95% confidence is used). Interestingly, the finite sample variance shrinks at the same time. Therefore for this particular DGP, incorporating many instruments not only leads to biased estimates, but also gives the illusion that the parameter is estimated precisely. With jackknife bias correction, there is much less empirical size distortion, and the empirical coverage rate remains well-controlled even with 200 instruments used in the first step. Moreover, the jackknife bias correction also (partially) restores the true variability of the estimator.

Although the focus here is on inference and, in particular, empirical coverage of associated testing procedures, it is also important to know how the bias correction will affect the Standard Deviation (sd) and the Mean Squared Error (MSE) of the point estimators. Recall that the model is correctly specified with 5 instruments, 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 instruments are included, however, the MSE increases rapidly without bias correction, while the MSE of the bias corrected estimator remains relatively stable. In particular, this finding is driven by a sharp reduction in bias that more than compensates the increase in variability of the estimator. A larger variance of the bias-corrected estimator is expected, as additional sampling variability is introduced by the bias correction. All in all, the bias-corrected estimator seems to be appealing not only for inference, but also for point estimation because it performs better in terms of MSE when the number of instruments is moderate or large.

In the Supplemental Appendix, we report results from two other data generating processes. In particular, we consider cases when (i) the true propensity score is non-linear and fundamentally

misspecified; and (ii) the true propensity score is non-linear and low-dimensional, and one employs basis expansion to approximate the true propensity score. The exact magnitude of the bias changes in different settings, but the same pattern emerges: as the number of included instruments/basis elements increases, the asymptotic distribution is no longer centered at the true parameter due to the bias uncovered in this paper (Theorem 1). Moreover, the jackknife continues to provide excellent bias correction (Theorem 2), and the bootstrap performs very well in approximating the finite sample distribution (Theorem 3).

7.2 Empirical Illustration

In this section we consider estimating the marginal returns to college education following the work of Carneiro, Heckman, and Vytlacil (2011, CHV hereafter) with MTE methods, employing the notation and assumptions imposed in Section 6.2. 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). 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 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 of college education. We make the same assumption as in CHV that the error terms are jointly independent of the covariates and the instruments. Then, $\tau_{\texttt{MTE}}(a|\mathbf{x}) = \partial\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}]/\partial a$ with

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

where $P_i = \mathbb{P}[T_i = 1|\mathbf{Z}_i]$ is the propensity score, and $\boldsymbol{\phi}$ is some fixed transformation. The covariates $\mathbf{X}_i$ include (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; and (iv) average local unemployment rate

and wage rate in 1991, and linear and square terms of work experience in 1991. For selection equation, the instruments $\mathbf{Z}_i$ include (i), (ii) and (iii) described earlier, as well as (v) the four raw instruments as well as their interactions with corrected AFQT score, education of mom and number of siblings. To make the functional form of the propensity score flexible, we also include interactions among the variables described in (i), and interactions between the cohort dummies and corrected AFQT score, education of mom and number of siblings. To conserve space, we leave summary statistics to the Supplemental Appendix.

We are employing the same covariates, instruments, and modeling assumptions as in CHV, but our estimation strategy is different than theirs. For the first step, the selection equation (propensity score) is estimated using a linear probability model with $k = 66$ as more interaction terms are included (which implies $k/\sqrt{n} = 1.58$), while CHV employ a Logit model with $k = 35$. Thus, our estimation approach reflects Assumption 1 in the sense that we assume away misspecification errors from using a flexible (high-dimensional) linear probability model, while CHV assume away misspecification errors from using a lower dimensional Logit model. For the second step, while the specification of $\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}]$ coincides, we estimate the partially linear model (that is, the $\boldsymbol{\phi}(a)$ component) using a flexible polynomial in P_i while CHV employ a kernel local polynomial approach with a bandwidth of about 0.30 over the support $[0, 1]$. To be specific, we implement the second step estimation by using least-squares regression with a fourth-order polynomial of the estimated propensity score $\boldsymbol{\phi}(\hat{P}_i) = [\hat{P}_i, \hat{P}_i^2, \hat{P}_i^3, \hat{P}_i^4]^\mathsf{T}$. Here the dimension of $\mathbf{X}_i$ is 23, so the second step model can be regarded as either "flexible" parametric or high-dimensional. In the latter case, the results reported in Section SA-4.3 of the Supplemental Appendix can be used, together with standard results from high-dimensional linear regression (see Cattaneo, Jansson, and Newey, 2017, 2018, and references therein), to show that a many covariate bias continues to be present in this setting, thereby justifying the usefulness of our fully automatic bias-correction and bootstrap-based methods. Finally, also in the Supplemental Appendix, we give results for other specifications of the selection and outcome equations.

We summarize the empirical findings in Figure 1, where we plot the estimated MTE evaluate at the sample average of $\mathbf{X}_i$. In the upper panel of this figure, we plot the estimated MTE together with 95% confidence intervals (solid and dashed blue line), using conventional two-step estimation methods (i.e., without bias correction and employing the standard normal approximation). These

empirical results are quite similar to those presented by CHV, both graphically and numerically. In particular, for individuals who are very likely to enroll in college, the per year return can be as high as 30%, while the return to college can also be as low as -20% for people who are very unlikely to enroll. Integrating the estimated MTE gives an estimator of the average treatment effect, which is roughly 9%.

The upper panel of Figure 1, also depicts the bias-corrected MTE estimator (dashed red line). The average treatment effect corresponding to the bias-corrected MTE is 8%, quite close to the previous estimate. On the other hand, the bias-corrected MTE curve has much steeper slope, implying a wider range of heterogeneity for returns to college education. This bias-corrected MTE curve lies close to the boundary of the confidence intervals constructed using the conventional two-step method, hinting at the possibility of a many instruments/covariate bias in the conventional estimate (blue line).

The lower panel of Figure 1 plots the bias corrected MTE estimator, together with the confidence intervals constructed using our proposed bootstrap-based method, which takes into account the extra variability introduced by bias correction. Not surprisingly, the new confidence intervals are wider than the conventional ones.

8 Conclusion

We studied the distributional properties of two-step estimators, and functionals thereof, when possibly many covariates are used to fit the first-step estimate (e.g., a propensity score, generated regressors or control functions). We show that overfitting in the first step estimation leads to a first-order bias in the distributional approximation of the two-step estimator. As a consequence, the limiting distribution is no longer centered at zero and usual inference procedures become invalid, possibly exhibiting severe empirical size distorsions in finite samples. We considered a few extensions of our basic framework and illustrated our generic results with several applications in treatment effect, program evaluation and other applied microeconomic settings. In particular, we presented new results for estimation and inference in the context of production function and marginal treatment effects estimation. The latter application, along with the one on local average response functions discussed in the Supplemental Appendix, give new results in the context of IV

models with treatment effect heterogeneity and many instruments, previously unavailable in the literature.

As a remedy for the many covariates bias we uncover, we develop bias correction methods using the jackknife. Importantly, this approach is data-driven and fully automatic, and does not require additional resampling beyond what would be needed to compute jackknife standard errors, which we show are also consistent in our setting even when many covariates are used. Therefore, implementation is straightforward and is available in any statistical computing platform. Furthermore, to improve finite sample inference after bias-correction, we also establish validity of an appropriately modified bootstrap for the jackknife-based bias-corrected Studentized statistic. We demonstrate the performance of our estimation and inference procedures in a comprehensive simulation study and an empirical illustration.

From a more general perspective, our main results give one additional contribution. They shed new light on the ultra-high-dimensional literature: one important implication is that typical sparsity assumptions imposed in that literature cannot be dropped in the context of non-linear models, since otherwise the effective number of *included* covariates will remain large after model selection, which in turn will lead to a non-vanishing first-order bias in the distributional approximation for the second-step estimator. It would be interesting to explore whether resampling methods are able to successfully remove this many *selected* or *included* covariates bias in ultra-high-dimensional settings, where model selection techniques are also used as a first-step estimation device.

Appendix A Omitted Regularity Conditions

The following set of technical conditions impose smoothness and bounded moments on various quantities. For simplicity, we denote by $\mathbf{m}_i = \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$, and make the following definitions: A random variable is said to be in BM_ℓ (bounded moments) if its ℓth moment is finite, and in BCM_ℓ (bounded conditional moments) if its ℓth conditional (on $\mathbf{z}_i$) moment is bounded uniformly by a finite constant. We also define the transformation

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

The same transformations are also applied to derivatives of $\mathbf{m}$.

Assumption A-1 (Smoothness and Bounded Moments)**.** Let $0 < \delta$, α, $C < \infty$ be some fixed constants.

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

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

(c) $\mathbf{m}$ is twice continuously differentiable in μ, with derivatives denoted by $\dot{\mathbf{m}}$ and $\ddot{\mathbf{m}}$, respectively.

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

(e) $\mathcal{H}^{\alpha,\delta}(\partial\dot{\mathbf{m}}_i/\partial\boldsymbol{\theta}) \in \mathsf{BM}_1$.

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231–263.

——— (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72(1), 1–19.

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70(1), 91–117.

Abadie, A., and M. D. Cattaneo (2018): "Econometric Methods for Program Evaluation," *Annual Review of Economics*, forthcoming.

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007): "Econometric tools for analyzing market outcomes," in *Handbook of Econometrics, Volume VI*, ed. by J. J. Heckman, and E. Leamer, pp. 4171–4276. Elsevier Science B.V., New York.

Alvarez, J., and M. Arellano (2003): "The Time Series and Cross-section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71(4), 1121–1159.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67(3), 499–527.

Bang, H., and J. M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, 85(1), 233–298.

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

Björklund, A., and R. Moffitt (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69(1), 42–49.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2017): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, forthcoming.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101(6), 2754–2781.

Cattaneo, M. D. (2010): "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155(2), 138–154.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2013): "Generalized Jackknife Estimators of Weighted Average Derivatives (With Comments and Rejoinder)," *Journal of the American Statistical Association*, 108(504), 1243–1256.

Cattaneo, M. D., and M. Jansson (2017): "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," Working paper, University of Michigan.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2017): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, forthcoming.

——— (2018): "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity," *Journal of the American Statistical Association*, forthcoming.

Chatterjee, S., and A. S. Hadi (1988): *Sensitivity Analysis in Linear Regression.* Wiley, New York.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volume VI*, ed. by J. J. Heckman, and E. Leamer, pp. 5549–5632. Elsevier Science B.V., New York.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models With Auxiliary Data," *Annals of Statistics*, 36(2), 808–843.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J. M. Robins (2017): "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, forthcoming.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, and W. K. Newey (2017): "Locally Robust Semiparametric Estimation," Working paper, MIT.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013): "On Robust Regression with High-Dimensional Predictors," *Proceedings of the National Academy of Sciences*, 110(36), 14557–14562.

Fan, J., J. Lv, and L. Qi (2011): "Sparse High-Dimensional Models in Economics," *Annual Review of Economics*, 3, 291–317.

Farrell, M. H. (2015): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," *Journal of Econometrics*, 189(1), 1–23.

Hahn, J., and W. K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Data Models," *Econometrica*, 72(4), 1295–1319.

Hahn, J., and G. Ridder (2013): "Asymptotic variance of semiparametric estimators with generated regressors," *Econometrica*, 81(1), 315–340.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

Heckman, J. J., and E. J. Vytlacil (2005): "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics, Volume VIB*, ed. by J. Heckman, and E. Leamer, pp. 5370–5468. Elsevier Science B.V.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Kline, P., and A. Santos (2012): "Higher order properties of the wild bootstrap under misspecification," *Journal of Econometrics*, 171(1), 54–70.

Li, C., and U. K. Müller (2017): "Linear Regression with Many Controls of Limited Explanatory Power," working paper, Princeton University.

Mammen, E. (1989): "Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap," *Annals of Statistics*, 17(1), 382–400.

——— (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Annals of Statistics*, 21, 255–285.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–82.

Newey, W. K., and D. L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Volume IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2111–2245. Elsevier Science B.V., New York.

Newey, W. K., and J. M. Robins (2017): "Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation," Working paper, MIT.

Olley, G. S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64(6), 1263–1297.

Vytlacil, E. J. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

Webb, M. D. (2014): "Reworking Wild Bootstrap Based Inference for Clustered Errors," *Working Paper*.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2 edn.

Wooldridge, J. M. (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50(2), 420–445.

Table 1. Simulation: Marginal Treatment Effects

(a) $n = 1000$

			Conventional					**Bias-Corrected**				
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	coverage	length	bias	sd	$\sqrt{\text{mse}}$	coverage	length
k												
5	0.00	0.16	0.14	4.72	4.73	0.95	18.51	−0.21	4.93	4.93	0.93	18.28
20	0.02	0.63	1.73	4.11	4.46	0.93	16.11	0.18	5.26	5.27	0.94	19.81
40	0.04	1.26	3.08	3.54	4.69	0.86	13.88	1.03	5.11	5.22	0.94	19.67
60	0.06	1.90	3.96	3.22	5.11	0.77	12.63	1.75	5.02	5.32	0.93	19.27
80	0.08	2.53	4.61	3.00	5.50	0.66	11.76	2.28	4.91	5.41	0.92	18.67
100	0.10	3.16	5.10	2.83	5.83	0.56	11.08	2.65	4.78	5.46	0.90	18.28
120	0.12	3.79	5.55	2.67	6.16	0.46	10.48	2.96	4.66	5.51	0.89	17.80
140	0.14	4.43	5.97	2.54	6.49	0.35	9.98	3.24	4.57	5.60	0.87	17.46
160	0.16	5.06	6.35	2.45	6.81	0.26	9.59	3.46	4.43	5.62	0.86	17.15
180	0.18	5.69	6.69	2.33	7.09	0.18	9.13	3.58	4.35	5.63	0.86	16.97
200	0.20	6.32	7.03	2.23	7.38	0.12	8.75	3.81	4.22	5.69	0.84	16.75

(b) $n = 2000$

			Conventional					**Bias-Corrected**				
	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	coverage	length	bias	sd	$\sqrt{\text{mse}}$	coverage	length
k												
5	0.00	0.11	0.13	4.85	4.85	0.95	19.00	−0.12	4.95	4.95	0.93	18.21
20	0.01	0.45	1.42	4.47	4.69	0.94	17.51	0.06	5.16	5.16	0.94	19.31
40	0.02	0.89	2.73	4.17	4.99	0.90	16.36	0.54	5.35	5.38	0.94	19.72
60	0.03	1.34	3.78	3.95	5.47	0.84	15.47	1.18	5.44	5.57	0.93	19.75
80	0.04	1.79	4.62	3.74	5.95	0.76	14.67	1.82	5.43	5.73	0.91	19.59
100	0.05	2.24	5.27	3.55	6.35	0.68	13.91	2.33	5.37	5.86	0.90	19.31
120	0.06	2.68	5.77	3.37	6.68	0.59	13.22	2.74	5.27	5.94	0.90	19.04
140	0.07	3.13	6.27	3.20	7.03	0.49	12.53	3.21	5.11	6.04	0.88	18.85
160	0.08	3.58	6.67	3.07	7.35	0.41	12.03	3.53	5.05	6.16	0.87	18.66
180	0.09	4.02	7.07	2.95	7.65	0.32	11.54	3.87	4.95	6.28	0.85	18.40
200	0.10	4.47	7.42	2.83	7.94	0.26	11.11	4.13	4.84	6.36	0.85	18.22

Notes. The marginal treatment effect is evaluated at $a = 0.5$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. Statistics are centered at the true value. $k = 5$ is the correctly specified model.
(i) k: number of instruments used for propensity score estimation;
(ii) bias: empirical bias (scaled by $\sqrt{n}$);
(iii) sd: empirical standard deviation (scaled by $\sqrt{n}$);
(iv) $\sqrt{\text{mse}}$: empirical root-MSE (scaled by $\sqrt{n}$);
(v) coverage: empirical coverage of a 95% confidence interval. Without bias correction, it is based on normal approximation and simulated sampling variability of the estimator (i.e. the oracle standard error). With bias correction, the test is based on the percentile-t method, where the bias-corrected and Studentized statistic is bootstrapped 500 times (Rademacher weights);
(vi) length: the average confidence interval length (scaled by $\sqrt{n}$).

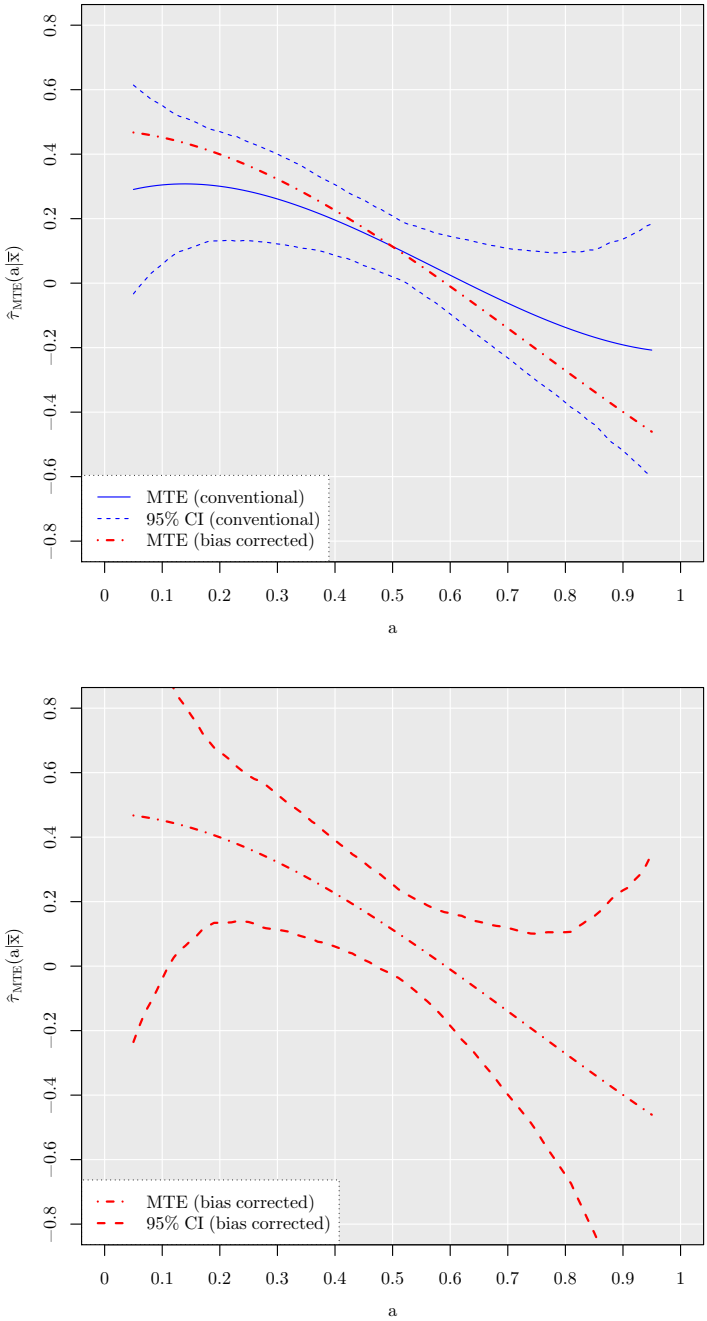

Figure 1. Marginal Treatment Effects

The marginal treatment effect, $\hat{\tau}_{\mathtt{MTE}}(a|\bar{\mathbf{X}})$, is evaluated at mean value of the covariates. Bootstrap is used to construct the confidence interval, with 500 repetitions.
Upper panel. Estimated MTE without bias correction (solid blue line), together with 95% confidence interval (dashed blue line). Also included is the bias-corrected MTE (dashed red line).
Lower panel. Bias-corrected MTE, together with 95% confidence interval, taking into account the effect of bias correction.

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

Matias D. Cattaneo† Michael Jansson‡ Xinwei Ma§

October 13, 2017

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 other treatment effect and policy evaluation examples not covered in the main paper, and reports detailed simulation evidence.

*This paper encompasses and supersedes our previous paper titled "Marginal Treatment Effects with Many Instruments", presented at the 2016 NBER summer meetings.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, Department of Statistics, University of Michigan.

Contents

SA-1 Setup, Notation and Basic Assumptions

This Supplemental Appendix is self-contained. We employ the same notation as in the main paper, but we reintroduce the setup and assumption to facilitate cross-referencing herein. Given a random sample $\{\mathbf{w}_i, \mu_i\}_{1\le i\le n}$, 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 $\{\mu_i\}_{1\le i\le n}$ are not directly observed. Instead, we assume the data available to the analyst consists of $\mathbf{w}_i = [\mathbf{y}_i^\mathsf{T}, r_i, \mathbf{z}_i^\mathsf{T}]^\mathsf{T}$, with $r_i \in \mathbb{R}$ and $\mathbf{z}_i \in \mathbb{R}^k$ always observed, and that the following first-step generated regressors condition holds:

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

The disturbance ε_i can be interpreted as a structural error term, or simply the error of 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:

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

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

where $\Theta \subset \mathbb{R}^{d_\theta}$ is the parameter space. We use $|\cdot|$ to denote the Euclidean norm, unless otherwise specified. To derive distributional properties, it is common to use the version:

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

We will use (E.5) primarily.

Next, we discuss basic notation and the main assumptions used throughout this Supplemental Appendix and, sometimes in stronger form, used in the main paper. Throughout, C is used to denote a generic (nonnegative and finite) constant, whose exact definition depends on the specific context. Recall that $|\cdot|$ denotes the Euclidean norm, and other norms will be defined at their

first appearance. 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 $\{b_n\}_{n\geq 1}$ (of numbers or random variables), we follow the convention that $a_n \precsim b_n$ if and only if $a_n \leq C_n \cdot b_n$, where $C_n = O(1)$. Similarly, we have $a_n \precsim_{\mathbb{P}} b_n$ if instead $C_n = O_{\mathbb{P}}(1)$

Assumption A.1 (Setup)**.**

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) $\hat{\mu}_i$ is uniformly consistent: $\max_{1\leq i\leq n} |\hat{\mu}_i - \mu_i| = o_{\mathbb{P}}(1)$.

A.1(4) $\hat{\boldsymbol{\theta}}$ satisfies (E.3), (E.5) and is tight.

A.1(5) The approximation error η_i in (E.2) satisfies $\max_{1\leq i\leq n} |\eta_i| = o_{\mathbb{P}}(1/\sqrt{n})$. ⌟

The next set of assumptions impose smoothness and bounded moments on various quantities. For simplicity, we denote by $\mathbf{m}_i = \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$, and make the following definitions: 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.

We also define the transformation

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

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

Assumption A.2 (Smoothness and Bounded Moments)**.** Let $0 < \delta$, α, $C < \infty$ be some fixed constants. And

A.2(1) $\mathcal{H}^{\alpha,\delta}(\mathbf{m}_i) \in \mathsf{BM}_1$.

A.2(2) $\mathbf{m}$ is continuously differentiable in $\boldsymbol{\theta}$ with $\mathcal{H}^{\alpha,\delta}(\partial \mathbf{m}_i/\partial \boldsymbol{\theta}) \in \mathsf{BM}_1$. Further, the matrix $\mathbf{M}_0 = \mathbb{E}\left[\partial \mathbf{m}_i/\partial \boldsymbol{\theta}\right]$ has full (column) rank d_θ.

A.2(3) $\mathbf{m}$ is twice continuously differentiable in μ, with derivatives denoted by $\dot{\mathbf{m}}$ and $\ddot{\mathbf{m}}$, respectively.

A.2(4) $\mathbf{m}_i$, $\dot{\mathbf{m}}_i$, $\ddot{\mathbf{m}}_i$, $\mathcal{H}^{\alpha,\delta}(\ddot{\mathbf{m}}_i)$, ε_i^2, $|\dot{\mathbf{m}}_i \varepsilon_i|$, $|\ddot{\mathbf{m}}_i|\varepsilon_i^2$, $|\mathcal{H}^{\alpha,\delta}(\ddot{\mathbf{m}})|\varepsilon_i^2 \in \mathsf{BCM}_2$. ⌟

SA-2 Primitive Conditions for First-Step Estimation

One critical assumption used in this paper is the uniform consistency of the first step estimate, A.1(3). We discuss primitive conditions in this section. Recall that the design matrix is $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]^\mathsf{T}$ and the projection matrix is $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-}\mathbf{Z}^\mathsf{T}$. First observe that

$$\max_{1\le i\le n} |\hat{\mu}_i - \mu_i| \le \max_{1\le i\le n} |\eta_i| + \max_{1\le i\le n} \left|\sum_j \pi_{ij}\left(\eta_j + \varepsilon_j\right)\right| \le \max_{1\le i\le n} |\eta_i| + \max_{1\le i\le n} \left|\sum_j \pi_{ij}\eta_j\right| + \max_{1\le i\le n} \left|\sum_j \pi_{ij}\varepsilon_j\right|,$$

where recall that π_{ij} denotes the (i,j) element of the projection matrix $\mathbf{\Pi}$. We study each of the terms above to show that $\max_{1\le i\le n} |\hat{\mu}_i - \mu_i| \to_{\mathbb{P}} 0$. We also discuss easy-to-verify primitive conditions for specific types of covariates $\mathbf{z}_i$.

SA-2.1 Linear Approximation Error

Using elementary inequalities, we obtain

$$\begin{aligned}
\max_{1\le i\le n} \left|\sum_j \pi_{ij}\eta_j\right| &\le \max_{1\le i\le n} |\eta_i| \max_{1\le i\le n} \sum_j |\pi_{ij}| = \max_{1\le i\le n} |\eta_i| \max_{1\le i\le n} n\frac{1}{n}\sum_j |\pi_{ij}| \\
&\le \max_{1\le i\le n} |\eta_i| n \max_{1\le i\le n} \sqrt{\frac{1}{n}\sum_j \pi_{ij}^2} = \max_{1\le i\le n} |\eta_i| \sqrt{n \max_{1\le i\le n} \sum_j \pi_{ij}^2} \\
&\le \max_{1\le i\le n} |\eta_i| \sqrt{n \max_{1\le i\le n} \pi_{ii}} \to_{\mathbb{P}} 0,
\end{aligned}$$

under the assumptions imposed in the paper. In particular, all it is needed at this point is

$$\max_{1\le i\le n} |\eta_i|^2 = o_{\mathbb{P}}\left(\frac{1}{n \max_{1\le i\le n} \pi_{ii}}\right),$$

which is implied by the assumption $\max_{1\le i\le n} |\eta_i| = o_{\mathbb{P}}(1/\sqrt{n})$ because $\max_{1\le i\le n} \pi_{ii} \le 1$.

SA-2.2 Residual Variability

Note that $\mathbb{E}[\varepsilon_i|\mathbf{Z}] = 0$, and this fact can be used to obtain sharp bounds. In particular, we illustrate here the case when the error term ε_i has (uniform) sub-Gaussian tail.

The ψ_p-norm of a random variable ε_i is defined as $\inf\{t \ge 0 : \ \mathbb{E}[\exp(\varepsilon_i^p/t^p)] \le 2\}$, and infimum taken over an empty set is understood to be $+\infty$. Then, Hoeffding's inequality gives

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

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. Then,

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

for any t. For the result above, the condition is enough to show that the conditional probability converges to 0 in probability. Since conditional probability is bounded, dominated convergence is used to show that the unconditional probability also converges to 0, which is the desired result. Therefore, the result follows by properties of the possibly many covariates $\mathbf{z}_i$ through the statistic $\max_{1\leq i\leq n} \pi_{ii}$.

The logic above can also be applied, albeit with different probability inequalities, to cases where the error term ε_i has thicker tails (e.g., of exponential or even polynomial decay).

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

The results above showed that the properties of the first-step estimator are determined by the behavior of the statistic $\max_{1\leq i\leq n} \pi_{ii}$, which in turn depends on the probabilistic properties of $\mathbf{z}_i$. In this section, we 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,

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

The upper bound can be used to give primitive conditions on different types of covariates $\mathbf{z}_i$. Here we focus on bounding $\max_{1\leq i\leq n} |\mathbf{z}_i|^2$ first, and then deducing the restrictions required on $\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z}/n)$.

Splines and other series expansions

If $\mathbf{z}_i$ is a Spline expansion, then

$$\max_{1\leq i\leq n} \pi_{ii} \precsim_{\mathbb{P}} \frac{1}{\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z}/n)} \cdot \frac{k}{n}.$$

In fact, under regularity conditions, $\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z}/n)$ will be bounded away from zero with probability approaching one, provided that $k\log(k)/n \to 0$. See, e.g., Belloni, Chernozhukov, Chetverikov, and Kato (2015) for a recent review and other examples of series expansions with similar properties.

Bounding through higher moments

A more general approach controls the higher moments of $|\mathbf{z}_i|^2$. For example, letting $\alpha > 2$,

$$\max_{1\leq i\leq n} |\mathbf{z}_i|^2 \precsim_{\mathbb{P}} n^{1/\alpha} \cdot \left(\mathbb{E}\left[|\mathbf{z}_i|^{2\alpha}\right]\right)^{1/\alpha},$$

which gives

$$\max_{1\le i\le n} \pi_{ii} \precsim_{\mathbb{P}} \frac{1}{\lambda_{\min}(\mathbf{Z}^\top \mathbf{Z}/n)} \cdot n^{\frac{1}{\alpha}-1} \cdot \left(\mathbb{E}\left[|\mathbf{z}_i|^{2\alpha}\right]\right)^{1/\alpha}.$$

Note that a crude bound for the above would be

$$\left(\mathbb{E}\left[|\mathbf{z}_i|^{2\alpha}\right]\right)^{1/\alpha} \le \sum_{\ell=1}^{k} \left(\mathbb{E}\left[|z_{\ell,i}|^{2\alpha}\right]\right)^{1/\alpha}.$$

Bounding by the tail of $\mathbf{z}_i$

A tighter bound can be obtained provided the tails of $z_{i,\ell}$ are well-controlled. For example, assume $\|z_{i,\ell}\|_{\psi_2} \le C$ for some C independent of ℓ or n. Then

$$\left\| |\mathbf{z}_i|^2 \right\|_{\psi_1} = \left\| \sum_{1\le \ell\le k} z_{i,\ell}^2 \right\|_{\psi_1} \le \sum_{1\le \ell\le k} \left\| z_{i,\ell}^2 \right\|_{\psi_1} = \sum_{1\le \ell\le k} \left\| z_{i,\ell} \right\|_{\psi_2}^2 \le kC^2.$$

The display above combines triangle inequality and the fact that a random variable is sub-Gaussian if and only if it is sub-exponential. Also we note that whether the random variable is centered does not affect the bound, since L^p norm is bounded by the ψ_2 norm up to a constant factor that depends only on p.

Then it implies $\max_{1\le i\le n} |\mathbf{z}_i|^2 \precsim_{\mathbb{P}} \log(n)\cdot k$. Then

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

Regression with dummy variables

In empirical work, it is not uncommon to encounter regression specifications including many dummy variables, such as year/region/group specific fixed effects and interactions among them. Here we illustrate how $\max_{1\le i\le n} \pi_{ii}$ can be controlled when many dummy variables are included. Let $\{z_{i,\ell}\}_{1\le \ell\le k}$ 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, hence it is very hard to control the ψ_2-norm of the vector. On the other hand, we still have the bound

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

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 $\underline{p}_n = \min_{1\le \ell\le k} p_{n,\ell}$. 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, $(\mathbf{Z}^\top\mathbf{Z}/n)^-$.) Therefore,

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

In fact, it is easy to see that, under the conditions given below, $\mathbb{P}[\min_{1\le\ell\le k} N_\ell > 0] \to 1$. To see its asymptotic order, consider (for some $\delta > 0$ and $t > 0$)

$$\begin{aligned}
&\mathbb{P}\left[\frac{n}{\min_{1\le\ell\le k} N_\ell} \ge \frac{t}{\underline{p}_n^{1+\delta}}\right] = \mathbb{P}\left[\min_{1\le\ell\le k} N_\ell \le \frac{1}{t} n\underline{p}_n^{1+\delta}\right] \\
&\le \sum_{\ell=1}^k \mathbb{P}\left[N_\ell \le \frac{1}{t} n\underline{p}_n^{1+\delta}\right] \le \sum_{\ell=1}^k \mathbb{P}\left[N_\ell \le \frac{1}{t} np_{n,\ell}^{1+\delta}\right] = \sum_{\ell=1}^k \mathbb{P}\left[N_\ell - np_{n,\ell} \le \frac{1}{t} np_{n,\ell}(p_{n,\ell}^\delta - t)\right] \\
&\le \sum_{\ell=1}^k \exp\left[-\frac{1}{2}\frac{t^{-2}n^2p_{n,\ell}^2(p_{n,\ell}^\delta - t)^2}{np_{n,\ell}(1-p_{n,\ell}) + \frac{1}{3}t^{-1}np_{n,\ell}|p_{n,\ell}^\delta - t|}\right] \qquad \text{(Bernstein's inequality)} \\
&= \sum_{\ell=1}^k \exp\left[-\frac{1}{2}\frac{t^{-2}np_{n,\ell}(p_{n,\ell}^\delta - t)^2}{1-p_{n,\ell} + \frac{1}{3}t^{-1}|p_{n,\ell}^\delta - t|}\right],
\end{aligned}$$

which goes to zero for any t, provided that (i) $\max_{1\le\ell\le k} p_{n,\ell}^\delta \to 0$; and (ii) $n\min_{1\le\ell\le k} p_{n,\ell}/\log(k) \to \infty$. Then, we have

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

One example is a balanced design, where for two constants C_1 and C_2, $C_1k^{-1} \le \min_{1\le\ell\le k} p_{n,\ell} \le \max_{1\le\ell\le k} p_{n,\ell} \le C_2k^{-1}$. Then, for $\alpha = 1/2$, and under the assumption $k = O(\sqrt{n})$,

$$\max_{1\le i\le n} \pi_{ii} = O_{\mathbb{P}}\left(\frac{k^{3/2}}{n}\right) = O_{\mathbb{P}}\left(n^{-1/4}\right).$$

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 illustrates the wide applicability of our results.

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 notation 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.1(3). 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, and using design balance can be a detour.

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 used in the first step, which is of rank k with probability approaching one. Then, $\sum_i \pi_{ii} = k$. 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; see Section SA-6. 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, hence otherwise it would not be possible to "delete one observation".

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

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

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

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 this 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 satisfies $p_{n,1} = c/n$ for some $c > 0$, while for the rest cells, $p_{n,\ell} = (1-c/n)/(k-1)$, $\ell \ge 2$. This captures the empirical relevant case where some cells may have very few observations. The problem here, however, is that N_1 follows Binomial distribution with $(n, c/n)$, hence has Poisson limiting distribution with mean c. And by the discussion in previous subsection, we have

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

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

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 $N_1 \ge C$, where $C \ge 2$ 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, note that

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

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)**.**
Assume A.1(1)–A.1(4) and A.2(1) hold. Then $\hat{\boldsymbol{\theta}}$ is consistent. That is, $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| = o_{\mathbb{P}}(1)$. ⌟

Next we consider large sample properties of the estimator, based on GMM framework. Assume $\hat{\boldsymbol{\theta}}$ is consistent and that $\boldsymbol{\theta}_0$ is an interior point of Θ, we have the following:

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

For notational convenience and later reference, let

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

With additional assumptions, we are able to better characterize through Taylor expansion:

Lemma SA.2 (First Linearization)**.**
Assume A.1 and A.2 hold, then

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) = \boldsymbol{\Sigma}_0 \left[\frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \right] \left(1 + o_{\mathbb{P}}(1) \right). \tag{E.6}$$

⌟

We can further apply Taylor expansion with respect to the first-step estimate $\hat{\mu}_i$, which implies the following:

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) && \text{(E.7)} \\
&+ \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \left(\hat{\mu}_i - \mu_i \right) && \text{(E.8)} \\
&+ \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \left(\hat{\mu}_i - \mu_i \right)^2 && \text{(E.9)} \\
&+ o_{\mathbb{P}}(1).
\end{aligned}$$

(E.7) is (part of) the usual influence function for parametric GMM problems. To handle (E.8), we first make the following decomposition:

$$\text{(E.8)} = \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_j \Big) + \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \eta_j \Big) - \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \eta_i.$$

Given the assumption that the approximation error η_i is of order $o_{\mathbb{P}}(1/\sqrt{n})$, the last two terms are easily shown to be of order $o_{\mathbb{P}}(1)$, hence we have the following lemma:

Lemma SA.3 (Term (E.8), 1)**.**
Assume A.1 and A.2 hold, then

$$(\text{E.8}) = \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_j \Big) + o_{\mathbb{P}}(1).$$

⌟

We upgrade Lemma SA.3 to the following result, which characterizes the (linear) bias contribution from over-fitting $\hat{\mu}_i$.

Lemma SA.4 (Term (E.8), 2)**.**
Assume A.1 and A.2 hold, then

$$(\text{E.8}) = \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0) \ |\mathbf{z}_j] \pi_{ij} \right) \cdot \varepsilon_i + \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + O_{\mathbb{P}} \left(\sqrt{\frac{k}{n}} \right) + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_{1,i} = \mathbb{E}\Big[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \cdot \varepsilon_i \ \Big| \mathbf{z}_i\Big]$. ⌟

Finally we give conditions under which it is possible to drop the double sum as well as the projection matrix in the variance component, which closes our discussion on Term (E.8).

Lemma SA.5 (Variance Simplification)**.**
Assume A.1 and A.2 hold. Further assume $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \ |\mathbf{z}_i]$ can be well approximated by $\mathbf{z}_i$ in mean squares:

$$\inf_{\boldsymbol{\Gamma}} \mathbb{E}\Big[\Big| \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) | \mathbf{z}_i] - \boldsymbol{\Gamma} \mathbf{z}_i \Big|^2 \Big] \to 0 \qquad \text{as } k \to \infty.$$

Then

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

⌟

Remark (Variance Contribution from the First Step)**.** For problems in which a nuisance parameter has to be estimated in a first step, usually the estimated nuisance parameter contributes to the asymptotic variance. This is well-documented in both the parametric and semi-parametric literature. In the current setting, the variance contribution from estimating μ_i is represented by

the term:

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

Note that when $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j,\mu_j,\boldsymbol{\theta}_0)\ |\mathbf{z}_j]=0$, estimating $\hat{\mu}_i$ will have no (first-order) impact on the variability of $\hat{\boldsymbol{\theta}}$. This is also quite intuitive: $\mathbf{m}$ is the estimating equation for $\boldsymbol{\theta}_0$, and $\dot{\mathbf{m}}$ represents the sensitivity of $\mathbf{m}$ with respect to the unobserved variable μ_i. See Newey (1994) for more discussions.

In the semi-parametric estimation literature, $\mathbf{z}_i$ is a series basis, hence the extra assumption in Lemma SA.5 is usually invoked to show that $\hat{\boldsymbol{\theta}}$ has an asymptotic linear representation. ⌟

Remark (Bias Contribution from Term (E.8))**.** When there are many covariates in the first step (i.e. when $\hat{\mu}_i$ is over-fitted), it also contributes to the asymptotic bias. Part of the bias is represented by $\mathbf{b}_{1,i}$. To see the intuition, note that a first order approximation gives $\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\approx\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)(\hat{\mu}_i-\mu_i)\approx\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)(\sum_j\pi_{ij}\varepsilon_j)$. Due to the conditional mean zero property of ε_j, the bias can be characterized by $\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_i\pi_{ii}$. Hence the bias (due to the linear contribution of $\hat{\mu}_i$) will be zero if (i) there is no residual variation in the sensitivity measure $\dot{\mathbf{m}}$: $\mathbb{V}[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\ |\mathbf{z}_i]=0$ almost surely; or (ii) the residual variation in the sensitivity measure $\dot{\mathbf{m}}$ is uncorrelated to the first step error term: $\mathbb{C}\mathrm{ov}[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0),\varepsilon_i\ |\mathbf{z}_i]=0$ almost surely.

Also note that if μ_i is estimated with a leave-out estimator (that is, using $\hat{\mu}_i^{(i)}$ instead of $\hat{\mu}_i$ in (E.3), the bias (due to the linear contribution of $\hat{\mu}_i$) will be zero. Later we will quantify the quadratic term, (E.9), and show that it also contributes to the asymptotic bias. Since the bias contribution from (E.9) is quadratic, using a leave-out estimator for μ_i will not be effective in removing such bias. ⌟

Next we consider the quadratic term (E.9) in the expansion. Again we first simplify, and show that the misspecification error η_i does not enter asymptotically.

Lemma SA.6 (Term (E.9), 1)**.**
Assume A.1 and A.2 hold, then

$$(\text{E.9})=\frac{1}{\sqrt{n}}\sum_i\frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\Big(\sum_j\pi_{ij}\varepsilon_j\Big)^2+o_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}+1\right).$$

⌟

Finally we upgrade Lemma SA.6 with conditional expectation and variance calculation.

Lemma SA.7 (Term (E.9), 2)**.**

Assume A.1 and A.2 hold, then

$$(\text{E.9}) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathbb{P}}(1),$$

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

Remark (Bias Contribution from Term (E.9))**.** In a previous remark, we characterize $\mathbf{b}_{1,i}$ as the bias which comes from the linear contribution of $\hat{\mu}_i$. In particular, the bias involving $\mathbf{b}_{1,i}$ will not appear if the leave-out version $\hat{\mu}_i^{(i)}$ is used. In this remark, we characterize the bias that arises due to the quadratic dependence of $\mathbf{m}$ on $\hat{\mu}_i$. Because of the quadratic nature, this bias represents the accumulated estimation error when $\hat{\mu}_i$ is over-fitted, and cannot be easily cured with simple method such as using the leave-out version $\hat{\mu}_i^{(i)}$.

Recall that $\mathbf{b}_{2,ij} = \frac{1}{2}\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \cdot \varepsilon_j^2 \,|\mathbf{z}_i, \mathbf{z}_j]$, and when $i \neq j$ (which is the main part of the bias), it reduces to a combination of $\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \,|\mathbf{z}_i]$ and $\mathbb{E}[\varepsilon_j^2 \,|\mathbf{z}_j]$. The latter is non-zero, hence to make $\mathbf{b}_{2,ij}$ zero, the only hope is that $\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \,|\mathbf{z}_i] = 0$. This fits the intuition quite well: over-fitting the first step does not make quadratic contribution if the estimating equation $\mathbf{m}$ is not sensitive to the second order. ⌟

The next proposition combines the previous lemmas, and gives an asymptotic representation of the estimator $\hat{\boldsymbol{\theta}}$.

Theorem SA.8 (Asymptotic Representation)**.**
Assume 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

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

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

⌟

Remark (Notation)**.** In this Supplemental Appendix, we use the notation $\boldsymbol{\mathcal{B}}$ 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 $\boldsymbol{\mathcal{B}}$ should be thought of 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: $\boldsymbol{\mathcal{B}} = \sqrt{n}\mathscr{B}$.

For the asymptotic representation, we use

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

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

Theorem SA.9 (Asymptotic Normality)**.**
Under the assumptions of the previous theorem,

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

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

In this section we consider extensions which arise in empirical applications.

SA-4.1 First Step: Multidimensional Case

Generalizing to vector-valued $\boldsymbol{\mu}_i$ is an easy and natural extension of our results, in the sense that all our proving strategies will go through and all results continue to hold under the same set of assumptions. On the other hand, the notation becomes more delicate/complicated since we keep not only the linear term in the expansion, but also the quadratic term. We first show how asymptotic representation of $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ changes when there are multiple unknowns estimated in the first step. In the next subsection, we illustrate the intuition and (partial) derivation when $\boldsymbol{\mu}_i$ is bivariate, and discuss the nature of the bias due to many covariates.

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,

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

with $\eta_{\ell i}$ being the approximation error and $\varepsilon_{\ell i}$ being the error from 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 $1 \le \ell \le d_\mu$. Note that when estimating the unknowns $\mu_{\ell i}$, we do allow different sets of covariates being used in the first step. Alternatively, one can think of $\mathbf{z}_i$ as a "long vector" which collects jointly the covariates used for estimating $\mu_{\ell i}$.

Both notation and assumptions have to be adjusted in this new setting. Let ℓ and ℓ' be two indices, we denote the derivatives of the estimating equation with respect to the unknowns as

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

Modified assumptions are postponed to the end of this section, and we first present the general asymptotic expansion including the influence function and the biases.

$$\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big) = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \mathbf{m}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \tag{E.10}$$

$$+ \sum_{\ell=1}^{d_\mu} \left[\frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \Big(\sum_j \mathbb{E}[\dot{\mathbf{m}}_\ell(\mathbf{w}_j, \boldsymbol{\mu}_j, \boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij} \Big) \cdot \varepsilon_{\ell i} \right] \tag{E.11}$$

$$+ \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} \tag{E.12}$$

$$+\frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\sum_{i,j}\mathbf{b}_{2,ij}\cdot\pi_{ij}^2+o_{\mathbb{P}}(1). \tag{E.13}$$

As before, (E.10) represents the influence function had $\boldsymbol{\mu}_i$ been observed, and (E.11) is the variance contribution from estimating $\boldsymbol{\mu}_i$. The bias terms (E.12) and (E.13) take similar form as before regarding how the projection matrix enters, with $\mathbf{b}_{1,i}$ and $\mathbf{b}_{2,ij}$ taking the following form

$$\mathbf{b}_{1,i}=\sum_{\ell=1}^{d_\mu}\mathbb{E}[\dot{\mathbf{m}}_\ell(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\varepsilon_{\ell i}|\mathbf{z}_i] \tag{E.14}$$

$$\mathbf{b}_{2,ij}=\sum_{\ell,\ell'=1}^{d_\mu}\frac{1}{2}\mathbb{E}[\ddot{\mathbf{m}}_{\ell\ell'}(\mathbf{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\theta}_0)\varepsilon_{\ell j}\varepsilon_{\ell' j}|\mathbf{z}_i,\mathbf{z}_j]. \tag{E.15}$$

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

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

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

then the analogue of Theorem SA.8 holds. Theorem SA.9 also holds provided that the variance does not vanish asymptotically.

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:

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

A.1(3) $\to \max_{1\le i\le n}|\hat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i|=o_{\mathbb{P}}(1)$.

A.1(5) $\to \max_{1\le i\le n}|\boldsymbol{\eta}_i|=o_{\mathbb{P}}(1/\sqrt{n})$.

A.2(3) $\to$ $\mathbf{m}$ is twice continuously differentiable in $\boldsymbol{\mu}$, with derivatives denoted by $\dot{\mathbf{m}}_\ell$ and $\ddot{\mathbf{m}}_{\ell\ell'}$, respectively.

A.2(4) $\to$ For all $1\le\ell,\ \ell'\le d_\mu$, $\mathbf{m}_i$, $\dot{\mathbf{m}}_{\ell,i}$, $\ddot{\mathbf{m}}_{\ell\ell',i}$, $\mathcal{H}^{\alpha,\delta}(\ddot{\mathbf{m}}_{\ell\ell',i})$, $|\boldsymbol{\varepsilon}_i|^2$, $|\dot{\mathbf{m}}_{\ell,i}||\boldsymbol{\varepsilon}_i|$, $|\ddot{\mathbf{m}}_{\ell\ell',i}||\boldsymbol{\varepsilon}_i|^2$, $|\mathcal{H}^{\alpha,\delta}(\ddot{\mathbf{m}}_{\ell\ell'})||\boldsymbol{\varepsilon}_i|^2\in\mathsf{BCM}_2$. ⌟

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

For illustration purpose, we consider $\boldsymbol{\mu}_i=[\mu_{1i},\mu_{2i}]^\mathsf{T}$ being bivariate. Again we start from the sample estimating equation, and linearize with respect to $\hat{\boldsymbol{\theta}}$, which yields (this is the analogue of

Lemma SA.2):

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

with

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\boldsymbol{\mu}}_i, \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) && \text{(E.16)}\\
&+ \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}_1(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_{1i} - \mu_{1i}\Big) && \text{(E.17)}\\
&+ \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}_2(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_{2i} - \mu_{2i}\Big) && \text{(E.18)}\\
&+ \frac{1}{\sqrt{n}} \sum_i \frac{1}{2}\ddot{\mathbf{m}}_{11}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_{1i} - \mu_{1i}\Big)^2 && \text{(E.19)}\\
&+ \frac{1}{\sqrt{n}} \sum_i \frac{1}{2}\ddot{\mathbf{m}}_{22}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_{2i} - \mu_{2i}\Big)^2 && \text{(E.20)}\\
&+ \frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_{1i} - \mu_{1i}\Big)\Big(\hat{\mu}_{2i} - \mu_{2i}\Big) && \text{(E.21)}\\
&+ o_{\mathbb{P}}(1).
\end{aligned}$$

As before, (E.16) is the influence function if $\boldsymbol{\mu}_i$ were observed; (E.17) and (E.18) represent the linear (leave-in) bias and variance contribution from estimating $\boldsymbol{\mu}_i$; and (E.19) and (E.21) are the quadratic biases.

In the same spirit as Lemma SA.3 and SA.4, we have the following result on Term (E.17) and (E.18).

$$\begin{aligned}
\text{(E.17)} &= \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}_1(\mathbf{w}_j, \boldsymbol{\mu}_j, \boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij} \right) \cdot \varepsilon_{1i} + \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,1,i}\pi_{ii} + o_{\mathbb{P}}(1),\\
\text{(E.18)} &= \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}_2(\mathbf{w}_j, \boldsymbol{\mu}_j, \boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij} \right) \cdot \varepsilon_{2i} + \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,2,i}\pi_{ii} + o_{\mathbb{P}}(1),
\end{aligned}$$

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 $\mathbf{b}_{1,2,i} = \mathbb{E}[\dot{\mathbf{m}}_2(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)\varepsilon_{2i}|\mathbf{z}_i]$. And as before, the above characterizes two linear contributions from estimating $\boldsymbol{\mu}_i$: variance contribution and leave-in biases. In addition, both the two bias terms are of order $O(k/\sqrt{n})$ by properties of the projection matrix.

Quadratic terms in the previous decomposition, at least (E.19) and (E.20), are handled by Lemma SA.6 and SA.7, yielding the following result.

$$\text{(E.19)} = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,11,ij}\pi_{ij}^2 + o_{\mathbb{P}}(1) \qquad \text{(E.20)} = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,22,ij}\pi_{ij}^2 + o_{\mathbb{P}}(1),$$

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 $\mathbf{b}_{2,22,ij} = \frac{1}{2}\mathbb{E}[\ddot{\mathbf{m}}_{22}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \cdot \varepsilon_{2j}^2 | \mathbf{z}_i, \mathbf{z}_j]$. And as before, by appealing to projection matrix properties, the two bias terms are of order $O(k/\sqrt{n})$.

For the new term (E.21), we first note that it has the following crude bound (which gives the order $k/\sqrt{n}$), a fact due to the Cauchy-Schwarz inequality:

$$
\begin{aligned}
|(\text{E.21})| &\leq \frac{1}{\sqrt{n}} \sum_i |\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)| \cdot |\hat{\mu}_{1i} - \mu_{1i}| \cdot |\hat{\mu}_{2i} - \mu_{2i}| \\
&\leq \sqrt{\frac{1}{\sqrt{n}} \sum_i |\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)|^2 \cdot |\hat{\mu}_{1i} - \mu_{1i}|^2} \sqrt{\frac{1}{\sqrt{n}} \sum_i |\ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0)|^2 \cdot |\hat{\mu}_{2i} - \mu_{2i}|^2} \precsim_{\mathbb{P}} \frac{k}{\sqrt{n}}.
\end{aligned}
$$

We would like to, however, have more precise characterization of (E.21), and the calculation follows the same strategy to prove Lemma SA.6 and SA.7. Conditional expectation calculation is given in the following, and we omit the conditional variance calculation. First we expand (E.21) as

$$
\begin{aligned}
(\text{E.21}) &= \frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \Big(\hat{\mu}_{1i} - \mu_{1i}\Big) \Big(\hat{\mu}_{2i} - \mu_{2i}\Big) \\
&= \frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_{12}(\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij}\varepsilon_{1j}\Big) \Big(\sum_j \pi_{ij}\varepsilon_{2j}\Big) + o_{\mathbb{P}}(1),
\end{aligned}
$$

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 $\sum_i \pi_{ii}^2/\sqrt{n}$. One example would be that $\mathbf{m}$ is linearly additive in the two unknowns μ_{1i} and μ_{2i}. Second, if correlation between the two error terms is zero (i.e. $\mathbb{E}[\varepsilon_{1j}\varepsilon_{2j} | \mathbf{z}_j] = 0$), bias contribution from this term is again of the 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. And 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

In this section we consider another generalization where the first step takes 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 remains to be the approximation error and ε_i is the residual from conditional expectation decomposition. Same as before, we assume $\boldsymbol{\beta}$ has dimension k which increases with the sample size, while $\boldsymbol{\gamma}$ has dimension d_γ which is *fixed*. (This is why we call it partially linear first step.) 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 modified. First it depends on $\boldsymbol{\gamma}$, which has to be estimated in the first step. Second μ_i enters the second step as a unknown. That is:

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

The real difficulty is not that $\boldsymbol{\gamma}$ enters the second step. Instead, the unknown that enters the second step now, μ_i, *is no longer a conditional expectation projection* (unless $\boldsymbol{\gamma}$ is known or $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are orthogonal). Fortunately we can rewrite the problem as:

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

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 above estimating equation, from which then $\hat{\boldsymbol{\theta}}$ is obtained. We show in this section that, despite a more general model is used, introducing the additional parameter $\boldsymbol{\gamma}$ in the first step only affects asymptotic variance, but not bias. In particular, our theory on asymptotic bias with many covariates entering the first step remains unchanged with the first step now taking a partially linear form.

Under very weak regularity conditions, one can show that $\hat{\boldsymbol{\gamma}}$ is $\sqrt{n}$-consistent for γ, and standard linearization technique shows that, after normalization and scaling, $\hat{\boldsymbol{\theta}}$ has the following representation:

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

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

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

with (we still use $\dot{\mathbf{m}}$ as the first partial derivative of $\mathbf{m}$ with respect to μ, and $\ddot{\mathbf{m}}$ the second derivative)

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

The next step is to further expand, which gives

$$\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big) = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0) \tag{E.22}$$

$$+ \boldsymbol{\Sigma}_0\boldsymbol{\Xi}_0\sqrt{n}\Big(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\Big) \tag{E.23}$$

$$+ \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\Big(\hat{\nu}_i - \nu_i\Big) \tag{E.24}$$

$$+ \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)\Big(\hat{\nu}_i - \nu_i\Big)^2 + o_\mathbb{P}(1). \tag{E.25}$$

We note that (E.22) is the influence function had both $\boldsymbol{\gamma}$ and μ_i been observed, and (E.23) is the total variance contribution from estimating $\boldsymbol{\gamma}$. (E.24) also has variance contribution since ν_i is estimated. Finally both (E.24) and (E.25) will lead to asymptotic bias under our many covariates assumption.

We first consider (E.24) and (E.25). Since $\hat{\nu}_i$ is constructed as linear projection, the same technique developed in Section SA-3 can be applied. To be more specific, 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 regularity conditions are enough to justify the following:

$$\text{(E.24)} = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_0)|\mathbf{z}_i]\pi_{ij}\right)\varepsilon_i + \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + o_\mathbb{P}(1),$$

$$\text{(E.25)} = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + o_\mathbb{P}(1),$$

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

Algebraically, the new term (E.23) can be rewritten as

$$\text{(E.23)} = \boldsymbol{\Sigma}_0\boldsymbol{\Xi}_0\Big(\frac{1}{n}\mathbf{Z}_1^\mathsf{T}\mathbf{Q}_2\mathbf{Z}_1\Big)^{-1}\frac{1}{\sqrt{n}}\sum_i\Big(\sum_j \mathbf{z}_{1j}q_{2ij}\Big)\varepsilon_i + o_\mathbb{P}(1),$$

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 (under very weak regularity conditions) 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) which establishes asymptotic normality of $\hat{\boldsymbol{\gamma}}$ in a much more general setting. We provide primitive conditions to justify Assumption 1–3 in Cattaneo, Jansson, and Newey (2018), which are sufficient to establish the desired result.

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

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

The same transformation is also applied to derivatives of the moment function. Then we make the following assumptions.

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

The first requirement is intuitive, which states that, after the high dimensional vector $\mathbf{z}_{2i}$ is partialed out, there is 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), 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. Note that our target parameter is $\hat{\boldsymbol{\theta}}$, which has the following 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

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

$$\bar{\boldsymbol{\Psi}}_1 = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\left[\sum_i \mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\gamma},\boldsymbol{\theta}_0)\right]$$

$$\bar{\boldsymbol{\Psi}}_2 = \boldsymbol{\Sigma}_0\boldsymbol{\Xi}_0\Big(\mathbf{Z}_1^\top\mathbf{Q}_2\mathbf{Z}_1\Big)^{-1}\frac{1}{\sqrt{n}}\sum_i\Big(\sum_j \mathbf{z}_{1j}q_{2ij}\Big)\varepsilon_i + \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\sum_i\left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\gamma},\boldsymbol{\theta}_0)|\mathbf{z}_i]\pi_{ij}\right)\varepsilon_i.$$

hence the condition we need is that, with probability approaching one, the minimum eigenvalue of the "overall" variance-covariance matrix being bounded away from zero. See Theorem SA.9.

SA-4.3 Second Step: Additional Many Covariates

It is very difficult, or even impossible, to extend Theorem SA.9 to a two-step problem where both steps are high dimensional in full generality. 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, to make the problem tractable we consider a setting where the first step estimate $\hat{\mu}_i$ is regarded as a generated regressor, which is plugged into a high-dimensional semi-linear regression problem. We show that new biases arise as a result of both the two steps having high dimension, and thus showing that the main conclusions of this 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 satisfying $k = O(\sqrt{n})$. Here 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). This is only for simplicity and can be relaxed, with the caveat that allowing for different high dimensional vectors in the two steps will make the final result much more cumbersome.

Given this setting, and assuming a non-linear least-squares model is considered, we can map this problem into a slight 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 $\mathbf{m}(\cdot)$. 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 since a possibly high-dimensional vector of covariates is now allowed for in the second step.

Some additional notation is needed to state the result. First, let $\mathbf{f} = \partial f / \partial \boldsymbol{\theta}$ be the derivative of f with respect to $\boldsymbol{\theta}$. Second, we use $q_{ij} = \mathbb{1}_{\{i=j\}} - \pi_{ij}$ to denote elements of the annihilator projection matrix $\mathbf{Q} = \mathbf{I} - \boldsymbol{\Pi}$. Third, we use $f_i = f(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0)$ and $\mathbf{f}_i = \mathbf{f}(\mathbf{x}_i, \mu_i, \boldsymbol{\theta}_0)$ to simplify exposition. Finally, the "dot" notation is reserved for partial derivatives with respect to μ as done

throughout this Supplemental Appendix.

Using a result similar to partitioned regression, that is, regressing out the high-dimensional vector $\mathbf{z}_i$, the estimator of interest $\hat{\boldsymbol{\theta}}$ is given by

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

where $\hat{\mu}_i$ is constructed from projecting r_i on $\mathbf{z}_i$ in a first step as done in our basic framework. Due to the presence of the high-dimensional vector in the second step, the asymptotic expansion becomes much more complicated. The first part of the following additional assumptions is employed to simplify the result; it is not essential to show the presence of asymptotic bias, though the formulas would be even more cumbersome without this approximation assumption.

Assumption (High Dimensional Second Step)**.**

A.HSS(1) The vector $\mathbf{z}_i$ can approximate (in the sense of Lemma SA.5) the following: $\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]$, $\mathbb{E}[\dot{\mathbf{f}}_i|\mathbf{z}_i]$, $\mathbb{E}[\mathbf{f}_i \dot{f}_i|\mathbf{z}_i]$ and $\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]\mathbb{E}[\dot{f}_i|\mathbf{z}_i]$.

A.HSS(2) The minimum eigenvalue of $\mathbb{E}[\mathbb{V}[\mathbf{f}_i|\mathbf{z}_i]]$ is bounded away from zero. ⌟

We leave the cumbersome computational details to Section SA-10.12. Under the assumptions of Theorem SA.8, A.HSS, and A.3(1), the following holds:

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

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 expansion shows that including high-dimensional covariates in the second-step estimation, when entering additively separably, leads to the presence of a many covariates bias of the same order as reported in our main Theorem SA.9. The actual form 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 the many covariates bias emerges even when μ_i enters the second step multiplicatively (i.e. f is linear in μ_i). See the last remark in this section for an example. 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 (2017, 2018), which is due to the intrinsic linearity of that setting.

We now discuss a few specialized examples to illustrate some implications of this extension.

Remark (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.9 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. ⌟

Remark (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 variance and bias remain essentially the same. ⌟

Remark (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 (following notation of the previous remark) $g_i=x_i\mu_i$, $\dot{g}_i=x_i$ and $\ddot{g}=0$. 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 $\Sigma_0=\mathbb{E}[\mu_i^2\mathbb{V}[x_i|\mathbf{z}_i]]^{-1}$

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

$$b_{4,ij\ell} = \theta_0 \mathbb{E}[x_i|\mathbf{z}_i]\mathbb{E}[x_\ell|\mathbf{z}_\ell]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j],$$

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

Due to the flexibility of the setting (E.1) and (E.2), our results cover a wide range of applications. In this section, we show that overfitting the first-step estimate gives a first order bias contribution for many estimators used in practice. In particular, here we consider several treatment effect and policy evaluation methods, as well as related problems, and give exact formulas of the bias and variance in each case. Within each example it is usually possible to understand the source of the bias better by combining the general bias formula with specific (identification) assumptions used for each estimator.

Remark (Notation)**.** To avoid notation conflict, we use uppercase letters to denote random variables, for example 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. Greek letters, functions and error terms may not be capitalized, unless necessary. ⌟

SA-5.1 Inverse Probability Weighting

We consider estimation via IPW in a general missing data problem. Our results apply immediately to treatment effect, data combination and measurement error settings, when a conditional independence assumption is imposed. Assume the parameter of interest is given by $\mathbb{E}[\mathbf{h}(\mathbf{Y}_i(1), \mathbf{X}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$, where $\mathbf{Y}_i(1)$ is subject to 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. Then, under the assumptions given below, the parameter $\boldsymbol{\theta}_0$ is identified 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. Without loss of generality, we assume $\mathbf{X}_i$ is a subvector of $\mathbf{Z}_i$ with fixed dimension. One example will be that $\mathbf{Z}_i$ is the series expansion of $\mathbf{X}_i$, hence conditioning on $\mathbf{Z}_i$ is the same as conditioning on $\mathbf{X}_i$.

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 $\mathbf{Y}_i(1)$. ⌟

To simplify the notation, 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 $\mathbf{Z}_i$.

To match notation so that previous results can be applied, we have

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

Applying Theorem SA.9, we have the following:

Proposition SA.10 (IPW)**.**
Under the assumptions of Theorem SA.9, and assume A.IPW holds. 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, $\inf_{\boldsymbol{\Gamma}} \mathbb{E}\left[\left|\mathbf{g}_i/P_i - \mathbf{Z}_i^\mathsf{T}\boldsymbol{\Gamma}\right|^2\right] = o(1)$, then

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

Remark (Bias)**.** 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. ⌟

Remark (Semi-parametric Efficiency)**.** From the above proposition, it becomes clear 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 good approximation power for $\mathbf{g}_i/P_i$. ⌟

We provide the following corollary, which specializes the previous conclusion to the special case where only an outcome variable is missing and the goal is to estimate its mean. Thus, we set $\mathbf{h}(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}_0) = Y_i - \theta_0$, and $\theta_0 = \mathbb{E}[Y_i(1)]$.

Corollary SA.11 (IPW: Estimating Mean of $y_i(1)$)**.**
Under the assumptions of Theorem SA.9, and assume A.IPW holds. 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, $\inf_{\boldsymbol{\gamma}} \mathbb{E}\left[\left|g_i/P_i - \mathbf{Z}_i^{\mathsf{T}}\boldsymbol{\gamma}\right|^2\right] = o(1)$, 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

Consider a simple setup where 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. We assume at time t_2 some individuals receive a treatment, and denote by $T_i = 1$ the treatment indicator. In a potential outcome framework, we can write the second period outcome as $Y_i(t_2) = Y_i(1,t_2)T_i + Y_i(0,t_2)(1-T_i)$, where $(Y_i(1,t_2), Y_i(0,t_2))$ are the potential outcomes in the second period receiving the treatment or not. 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 assumption used to achieve identification in this context is the so-called "parallel trends" assumption. Abadie (2005) relaxes that 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 $0 < C < 1/2$, such that $C \leq \mathbb{P}[T_i = 1|\mathbf{X}_i] \leq 1 - C$. ⌟

With Assumptions A.DiD(1) and A.DiD(2), and regularity conditions such as bounded moments, it is easily to show that

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

where $P_i = \mathbb{P}[T_i = 1|\mathbf{X}_i]$ is the propensity score and the expression is identified from the marginal distribution of the observed quantities. To see this, by conditioning on $\mathbf{X}_i$ and separating into two scenarios $T_i = 0, 1$, we have

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

Then, using Assumption A.DiD(1), we obtain

$$\begin{aligned}
&\mathbb{E}\left\{\frac{P_i}{\mathbb{P}[T_i = 1]}\left\{\mathbb{E}\left[\Big(Y_i(1,t_2) - Y_i(t_1)\Big)\Big|\mathbf{X}_i, T_i = 1\right] - \mathbb{E}\left[\Big(Y_i(0,t_2) - Y_i(t_1)\Big)\Big|\mathbf{X}_i, T_i = 1\right]\right\}\right\} \\
&= \mathbb{E}\left\{\frac{P_i}{\mathbb{P}[T_i = 1]}\mathbb{E}\left[\Big(Y_i(1,t_2) - Y_i(0,t_2)\Big)\Big|\mathbf{X}_i, T_i = 1\right]\right\} \\
&= \int \frac{P_i}{\mathbb{P}[T_i = 1]}\mathbb{E}\left[\Big(Y_i(1,t_2) - Y_i(0,t_2)\Big)\Big|\mathbf{X}_i, T_i = 1\right]\mathbb{P}(\mathrm{d}\mathbf{X}_i) \\
&= \int \mathbb{E}\left[\Big(Y_i(1,t_2) - Y_i(0,t_2)\Big)\Big|\mathbf{X}_i, T_i = 1\right]\mathbb{P}(\mathrm{d}\mathbf{X}_i|T_i = 1) \\
&= \mathbb{E}\left\{\mathbb{E}\left[\Big(Y_i(1,t_2) - Y_i(0,t_2)\Big)\Big|\mathbf{X}_i, T_i = 1\right]\Big|T_i = 1\right\} = \theta_0.
\end{aligned}$$

To estimate θ_0, it suffices to use a two-step procedure, where first the propensity score is estimated. Abadie (2005) proposes to estimate P_i by regressing T_i on a series expansion of $\mathbf{X}_i$, which is coved by our framework. The estimating equation is

$$\begin{aligned}
0 &= \mathbb{E}\left[\frac{T_i - P_i}{1 - P_i}\Big(Y_i(t_2) - Y_i(t_1)\Big) - T_i\theta_0\right] \\
\Rightarrow \qquad 0 &= \frac{1}{n}\sum_i \left[\frac{T_i - \hat{P}_i}{1 - \hat{P}_i}\Big(Y_i(t_2) - Y_i(t_1)\Big) - T_i\hat{\theta}\right] \\
\Leftrightarrow \qquad \hat{\theta} &= \sum_i \left[\frac{T_i - \hat{P}_i}{1 - \hat{P}_i}\Big(Y_i(t_2) - Y_i(t_1)\Big)\right] \Big/ \sum_i T_i.
\end{aligned}$$

To match notation, we have

$$\mathbf{w}_i = [T_i, Y_i(\cdot)]^\mathsf{T},\ r_i = T_i,\ \mu_i = P_i,\ \mathbf{z}_i = \text{series expansion of } \mathbf{X}_i$$

$$\mathbf{m}(\mathbf{w}_i, \mu_i, \theta) = \frac{T_i - P_i}{1 - P_i}\Big(Y_i(t_2) - Y_i(t_1)\Big) - T_i\theta.$$

We have the following result, which is Theorem SA.8 specialized to this model.

Proposition SA.12 (DiD)**.**
Under the assumptions of Theorem SA.9, and assume A.DiD holds. 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, $\inf_{\boldsymbol{\gamma}} \mathbb{E}\left[\left|\frac{1}{1-P_i} g_i - \mathbf{Z}_i^{\mathsf{T}}\boldsymbol{\gamma}\right|^2\right] = o(1)$, where $\mathbf{Z}_i$ is the series expansion of $\mathbf{X}_i$, 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

This is one of the examples analyzed in the main paper. Here we give further details and related regularity conditions, not discussed in the main paper to conserve space. To re-introduce the results of Abadie (2003), we need the notion of potential outcomes. Let $D_i \in \{0, 1\}$ be the (binary) instrumental variable, then the treatment indicator $T_i \in \{0, 1\}$ is a combination of two potential outcomes, $T_i = D_i T_i(1) + (1 - D_i)T_i(0)$. And let Y_i be the observed outcome, it will be the combination of four potential outcomes, corresponding to different values taken by the instrumental variable and the treatment indicator: $Y_i = T_i D_i Y_i(1, 1) + T_i(1 - D_i)Y_i(1, 0) + (1 - T_i)D_i Y_i(0, 1) + (1 - T_i)(1 - D_i)Y_i(0, 0)$. That is, the first argument of $Y_i(\cdot, \cdot)$ corresponds to the treatment status, and the second corresponds to the value of the instrument. For identification, we rely on the following assumptions, where $\mathbf{Z}_i$ are some 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 $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$ 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.

Let $g(Y_i, T_i, \mathbf{Z}_i)$ be a measurable function with finite first moment, Abadie (2003) showed the following identification result:

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

and as before, $P_i = \mathbb{E}[D_i|\mathbf{Z}_i] = \mathbb{P}[D_i = 1|\mathbf{Z}_i]$. To see the usefulness of the above result, note that in practice, it is not possible to identify the compliers. Although one can identify the local average treatment effect, it is not obvious how it should be connected to other treatment parameters, such as the average treatment effect or the treatment effect on the treated, since the compliers can have very different characteristics. With the above method, any characteristics depending only on the joint distribution of Y_i, T_i and $\mathbf{Z}_i$ can be identified for compliers, hence it is possible to give summary statistics for the compliers.

The previous result has another important empirical application: it allows one to fit a model for the outcome variable Y_i with prespecified functional form. For simplicity, let $\mathbf{X}_i$ be a subvector of $\mathbf{Z}_i$ with fixed dimension d_x, and assume one is interested in the conditional expectation function $\mathbb{E}[Y_i|T_i, \mathbf{X}_i, T_i(1) > T_i(0)]$. In general it will be nonlinear and can only be identified nonparametrically. To avoid the curse of dimensionality and ensure the result being interpretable, it is common to fit a best linear approximation to the conditional expectation function. That is,

$$\boldsymbol{\theta}_0 = \begin{bmatrix} \gamma_0 \\ \boldsymbol{\phi}_0 \end{bmatrix} = \arg\min_{\gamma_0, \boldsymbol{\phi}_0} \mathbb{E}\Big[\Big(Y_i - T_i\gamma_0 - \mathbf{X}_i^\intercal \boldsymbol{\phi}_0\Big)^2 \Big|T_i(1) > T_i(0)\Big].$$

We will take a slightly more general approach. Let e be a known link function which is smooth, we define the parameter of interest to be

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \mathbb{E}\Big[\Big(Y_i - e(\mathbf{X}_i, T_i, \boldsymbol{\theta})\Big)^2 \Big|T_i(1) > T_i(0)\Big],$$

where the linear model becomes a special case. For simplicity, we assume $\boldsymbol{\theta}_0$ is identified by the following moment condition,

$$\mathbb{E}\Big[\frac{\partial e}{\partial \boldsymbol{\theta}}(\mathbf{X}_i, T_i, \boldsymbol{\theta})\Big(Y_i - e(\mathbf{X}_i, T_i, \boldsymbol{\theta})\Big)\Big|T_i(1) > T_i(0)\Big] = \mathbf{0} \qquad \Leftrightarrow \qquad \boldsymbol{\theta} = \boldsymbol{\theta}_0.$$

Despite the simple linear functional form, it is not clear how $\boldsymbol{\theta}_0$ can be identified, since it requires conditioning on the compliers. With the results of Abadie (2003), one has

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

which will be the (population) estimating equation considered in this subsection. Note that the RHS depends only on observed variables.

To match our previous notations, we have

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

Essentially, the above is a weighted nonlinear least squares problem where the weights have to be estimated. (Also note that the weights can take negative values.) We have the following result, which is essentially Proposition SA.8 specialized to the current context.

Proposition SA.13 (LARF)**.**
Under the assumptions of Theorem SA.9, and assume A.LARF holds. 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

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

and

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

$$e_i(\boldsymbol{\theta}) = e(\mathbf{X}_i, T_i, \boldsymbol{\theta})$$
$$\boldsymbol{\Sigma}_0 = \left(-\mathbb{E}\left[\kappa_i \cdot \left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) - \frac{\partial}{\partial\boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\frac{\partial}{\partial\boldsymbol{\theta}^\intercal} e_i(\boldsymbol{\theta}_0)\right]\right]\right)^{-1}.$$

⌟

Remark (Bias and Variance: Beyond Compliers)**.** The bias and the asymptotic representation in the current context take complicated forms, and it is not obvious how those terms can be interpreted intuitively. On the other hand, it is clear that both the bias and the variance (more specificaly, the part of the variance contributed by the first step) are related to quantities averaged over *always-takes* and *never-takers*. And when there are no always-takers or never-takers (hence by Assumption A.LARF(4), there are only compliers), the bias will be zero, and the estimated first step does not contribute to the variance. This is not surprising, since when there are only compliers, it becomes a text-book linear regression problem, and the weights κ_i are identically 1. ⌟

SA-5.4 Marginal Treatment Effect

This is another example analyzed in the main paper, which can be interpreted as a generalization of the local average response function when the instrument is not binary. The marginal treatment effect (MTE) was proposed by Björklund and Moffitt (1987), and later developed by Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006). When identified, it is a key parameter to understand the treatment effect heterogeneity, and can also be used to construct other treatment parameters.

Assume one has a random sample and let Y_i be the outcome variable. We adopt the potential outcome framework, hence $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$, where T_i is the treatment indicator. Given some covariates $\mathbf{X}_i \in \mathbb{R}^{d_x}$, we decompose the potential outcomes into the conditional expectations and errors,

$$Y_i(0) = g_0(\mathbf{X}_i) + U_{0i}, \qquad Y_i(1) = g_1(\mathbf{X}_i) + U_{1i}.$$

The selection rule is determined by

$$T_i = \mathbb{1}\Big[P_i \geq V_i\Big].$$

and we impose the assumption that v_i is uniformly distributed in $[0, 1]$ conditional on $\mathbf{X}_i$. The marginal treatment effect (MTE) is defined as $\tau(a|\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0)|V_i = a, \mathbf{X}_i = \mathbf{x}]$. Heckman and Vytlacil (2005) provides a comprehensive introduction to the interpretation of the marginal treatment effect and its empirical importance in program evaluation. Assumptions A.MTE(1) and A.MTE(2) allow one to identify the MTE based on the following local instrumental variable

approach:

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

To avoid the curse of dimensionality, we adopt the following parametrization:

$$\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}] = e(\mathbf{x}, a, \boldsymbol{\theta}_0), \qquad \tau(a|\mathbf{x}) = \frac{\partial}{\partial p}e(\mathbf{x}, p, \boldsymbol{\theta}_0)\Big|_{p=a},$$

where e is known function up to the unknown parameter $\boldsymbol{\theta}_0$. Another complication is that P_i (a.k.a. the propensity score), is not observed, and has to be estimated in a first step. To make it more concrete, consider

$$T_i = P_i + \varepsilon_i, \qquad P_i = \mathbb{E}[T_i|\mathbf{Z}_i].$$

where we assume there are some instruments observed by the analyst, denoted by $\mathbf{Z}_i$, which can be used as the covariates to estimate the propensity score. Again we defer the identification assumptions. Now the problem can be reframed into the general model, and we state it as a Z-estimation:

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

Or equivalently, $\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}e(\mathbf{X}_i, P_i, \boldsymbol{\theta})\Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta})\Big)$. Usually the parameter of interest is not $\boldsymbol{\theta}_0$, but the MTE curve, or a weighted average of the MTE. Note that once we establish the asymptotic theory of $\hat{\boldsymbol{\theta}}$, we can rely on the delta method to infer the properties of the estimate MTE curve or its weighted average. Therefore we will devote this subsection to the theories of $\hat{\boldsymbol{\theta}}$. Following are the identification assumptions:

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

To save the notation, we use the following:

$$e_i(\boldsymbol{\theta}) = e(\mathbf{X}_i, P_i, \boldsymbol{\theta}), \qquad \dot{e}_i(\boldsymbol{\theta}) = \frac{\partial}{\partial p}e(\mathbf{X}_i, p, \boldsymbol{\theta})\Big|_{p=P_i}, \qquad \ddot{e}_i(\boldsymbol{\theta}) = \frac{\partial^2}{\partial p^2}e(\mathbf{X}_i, p, \boldsymbol{\theta})\Big|_{p=P_i}.$$

And to match notation, let

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

Then one has

Proposition SA.14 (MTE)**.**
Under the assumptions of Theorem SA.9, and assume A.MTE 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

$$\begin{aligned}
\boldsymbol{\mathcal{B}} &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left[\sum_i \mathbf{b}_{1,i}\cdot\pi_{ii} + \sum_{i,j}\mathbf{b}_{2,ij}\cdot\pi_{ij}^2\right] \\
\bar{\boldsymbol{\Psi}}_1 &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}} e_i(\boldsymbol{\theta}_0)\Big(Y_i - e_i(\boldsymbol{\theta}_0)\Big) \\
\bar{\boldsymbol{\Psi}}_2 &= \boldsymbol{\Sigma}_0 \frac{-1}{\sqrt{n}}\sum_i \left(\sum_j \frac{\partial}{\partial\boldsymbol{\theta}} e_j(\boldsymbol{\theta}_0)\cdot\dot{e}_j(\boldsymbol{\theta}_0)\cdot\pi_{ij}\right)\varepsilon_i,
\end{aligned}$$

and

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

⌟

Remark (Bias)**.** To understand the implications of the above theorem, we consider the bias terms. Note that $\mathbf{b}_{1,i}$ essentially captures treatment effect heterogeneity (in the outcome equation) and self-selection. To make it zero, one needs to assume there is no heterogeneous treatment effect and the agents do not act on idiosyncratic characteristics that are unobservable to the analyst. For the second bias term $\mathbf{b}_{2,ij}$, note that $\dot{e}_i(\boldsymbol{\theta}_0)$ is simply the MTE and $\ddot{e}_i(\boldsymbol{\theta}_0)$ is its curvature. Hence the second bias is related not only to treatment effect heterogeneity captured through the shape of the MTE, but also to the "level" of the MTE. As as consequence, the many instruments bias will not be zero unless there is no self-selection and the treatment effect is homogeneous and zero. ⌟

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. Usually the control function approach requires a first-step estimation and excluded instruments. For a recent review of the control function approach see Wooldridge (2015).

Due to its population in applied work, we will focus on the 2SLS estimator in this subsection, and frame it as a linear control function approach (we discuss the non-linear case further below). 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, Angrist, and Krueger (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

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

As before, we denote $\mu_i = \mathbf{Z}_i^\mathsf{T}\boldsymbol{\beta}$ (note that we simply assumed there is no "misspecification error"). 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 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 (hence, 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.1(3) is used to justify the second line. An alternative estimator is the JIVE proposed by Imbens, Angrist, and Krueger (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 should not be 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 will consider the control function approach in a nonlinear 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):

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

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:

$$
\begin{aligned}
\mathbb{E}[Y_i | X_i, \mathbf{Z}_i] &= \mathbb{P}\left[Y_i = 1 | X_i, \mathbf{Z}_i\right] = \mathbb{P}\left[u_i \leq X_i \delta_0 | X_i, \varepsilon_i\right] \\
&= \Phi\left(X_i \tilde{\delta}_0 - (X_i - \mathbf{Z}_i^{\mathsf{T}} \boldsymbol{\beta}) \gamma_0\right),
\end{aligned}
$$

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

$$
\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)] = \mathbb{E}\left[\mathbf{X}_i L'(\mathbf{X}_i^{\mathsf{T}} \boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^{\mathsf{T}} \boldsymbol{\theta}_0)\Big)\right] = \mathbf{0}, \tag{E.26}
$$

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 degenerate and not 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.26), with some known link function L such that $\mathbb{E}[Y_i|X_i, \mathbf{Z}_i] = L(\mathbf{X}_i^\top \boldsymbol{\theta}_0)$.

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

Remark. 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 one takes (E.26) as the estimating equation and $\boldsymbol{\theta}_0$ defined thereof as the parameter of interest. Of course, by dropping those assumptions, $\boldsymbol{\theta}_0$ no longer has structural interpretation. We maintain those assumptions to simplify the formula of the bias and variance. ⌟

Proposition SA.15 (Control Function)**.**
Under the assumptions of Theorem SA.9, and assume A.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

$$\begin{aligned}
\boldsymbol{\mathcal{B}} &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left[\sum_i \mathbf{b}_{1,i}\cdot \pi_{ii} + \sum_{i,j} \mathbf{b}_{2,ij}\cdot \pi_{ij}^2\right] \\
\bar{\boldsymbol{\Psi}}_1 &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\sum_i \mathbf{X}_i L'(\mathbf{X}_i^\top \boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top \boldsymbol{\theta}_0)\Big) \\
\bar{\boldsymbol{\Psi}}_2 &= \boldsymbol{\Sigma}_0 \frac{-1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}[\gamma_0 \mathbf{X}_j L'(\mathbf{X}_j^\top \boldsymbol{\theta}_0)^2 | \mathbf{Z}_j]\cdot \pi_{ij}\right)\varepsilon_i,
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{b}_{1,i} &= -\mathbb{E}[\gamma_0 \mathbf{X}_i L'(\mathbf{X}_i^\top \boldsymbol{\theta}_0)^2 \varepsilon_i | \mathbf{Z}_i] \\
\mathbf{b}_{2,ij} &= \mathbb{E}\left[\frac{\gamma_0}{2}\mathbf{e}_2 L'(\mathbf{X}_i^\top \boldsymbol{\theta}_0)^2 \varepsilon_j^2 - \frac{\gamma_0^2}{2}\mathbf{X}_i L''(\mathbf{X}_i^\top \boldsymbol{\theta}_0) L'(\mathbf{X}_i^\top \boldsymbol{\theta}_0)\varepsilon_j^2 - \gamma_0^2 \mathbf{X}_i L''(\mathbf{X}_i^\top \boldsymbol{\theta}_0)\varepsilon_j^2 \,\middle|\, \mathbf{Z}_i, \mathbf{Z}_j\right] \\
\boldsymbol{\Sigma}_0 &= \left(\mathbb{E}\left[\mathbf{X}_i \mathbf{X}_i^\top L'(\mathbf{X}_i^\top \boldsymbol{\theta}_0)^2\right]\right)^{-1}.
\end{aligned}$$

Remark (Exogenous X_i)**.** 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.) ⌟

Remark (JIVE). 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 fact in non-linear models. ⌟

SA-5.7 Production Function Estimation

In this section we consider the problem of estimating production functions, following the setup of Olley and Pakes (1996). Assume the production function takes Cobb-Douglas form, with four factors entering: labor input $L_{i,t}$, capital input $K_{i,t}$, the effect of aging on production $A_{i,t}$, and a productivity factor $W_{i,t}$. Hence denote by $Y_{i,t}$ the (log) production of firm i at time t, it is given by

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

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

Investment decision $I_{i,t}$ is based on the productivity factor, hence under some identification assumptions, it is possible to invert the relation and write $W_{i,t}$ as

$$W_{i,t} = h_t(I_{i,t}, K_{i,t}, A_{i,t}),$$

for some unknown and time-dependent function h_t. Therefore the output $Y_{i,t}$ becomes

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

In what follows, we use $\phi_{i,t}$ to denote $\phi_t(I_{i,t}, K_{i,t}, A_{i,t})$ whenever there is no confusion on the time index. Note that the above display can be used to estimate the labor share β_L, but not β_K or β_A due to the presence of the unknown function h_t.

Finally by taking conditional expectation of $W_{i,t+1}$ on time-t information, and that firm i survives at $t+1$, we have the following (note the difference in time indices):

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

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

$W_{i,t+1}$, and $P_{i,t}$ is the survival rate, defined as

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

and $\chi_{i,t+1}$ is the indicator of firm in business. Before moving on, we make a remark on the two error terms. The first one, $U_{i,t+1}$, is orthogonal with time-$t+1$ information. On the other hand, $V_{i,t+1}$ is obtained from expectation conditional on time-t information, hence it is *not* orthogonal to information at $t+1$. For this reason, it is can correlate 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".

Now we explain how the three parameters, β_L, β_K and β_A are estimated. For simplicity, we assume there are only two time periods $t \in \{t_1, t_2\}$. Then 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. That is, we regress 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 $\hat{\phi}_{i,t_1}$. In terms of notation, we have

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

Another first step is needed to estimate P_{i,t_1}. This is done by regressing/projecting 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 $\hat{P}_{i,t_1}$. We have the following to match notation:

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

Finally, β_K and β_A are estimated by (we assume the function g is known up to a finite dimensional nuisance parameter $\boldsymbol{\lambda}$)

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

which is a standard nonlinear least squares problem. Note that 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.

Transforming into this form, it becomes clear that all our results apply to this example, with two

minor generalizations proposed in Section 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 $\boldsymbol{\theta} = [\beta_K, \beta_A, \boldsymbol{\lambda}^\mathsf{T}]^\mathsf{T}$, which is solved from the sample moment condition

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

with

$$\mathbf{w}_i = [Y_{i,t_2}, L_{i,t_2}, K_{i,t_2}, A_{i,t_2}, L_{i,t_1}, K_{i,t_1}, A_{i,t_1}],\ \hat{\gamma} = \hat{\beta}_L.$$

The above corresponds to moment conditions for β_K, β_L and the nuisance parameter $\boldsymbol{\lambda}$. Also we denote by g_ℓ the derivative of g with respect to its ℓ-th argument. Same analogy is used for higher order derivatives.

Proposition SA.16 (Production Function)**.**
Under the assumptions of Theorem SA.9, and assume A.PL holds and $\sum_i \pi_{ii}^2 = o_\mathbb{P}(k)$. 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

$$\begin{aligned}
\boldsymbol{\mathcal{B}} &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \left[\sum_i \Big(\mathbf{b}_{1,1,i} + \mathbf{b}_{1,2,i}\Big)\pi_{ii} + \sum_{i,j} \Big(\mathbf{b}_{2,11,ij} + \mathbf{b}_{2,22,ij} + \mathbf{b}_{2,12,ij}\Big)\pi_{ij}^2 \right] \\
\bar{\boldsymbol{\Psi}}_1 &= \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} \Big(V_{i,t_2} + U_{i,t_2}\Big) \\
\bar{\boldsymbol{\Psi}}_2 &= -\frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_i \left\{ \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{2,i,t_1} U_{i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{1,i,t_1}\Big(\chi_{i,t_2} - P_{i,t_1}\Big) \right\} \\
&\quad + \frac{1}{\sqrt{n}} \frac{1}{\mathbb{EV}[L_{i,t_1}|(I,K,A)_{i,t_1}]} \boldsymbol{\Sigma}_0 \boldsymbol{\Xi}_0 \sum_i \Big(L_{i,t_1} - \mathbb{E}[L_{i,t_1}|(I,K,A)_{i,t_1}]\Big) U_{i,t_1}.
\end{aligned}$$

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 formulae for $\mathbf{\Sigma}_0$ and $\mathbf{\Xi}_0$ (defined formally in Section SA-4.2), since they are quite long and yet can be derived easily from their definitions. Also we made the additional assumption that $\sum_i \pi_{ii}^2 = o_{\mathbb{P}}(k)$ to simplify the bias formula. Note that the previous result remains true without this assumption, albeit the biases becomes even more cumbersome.

Remark (Bias)**.** Some bias terms can be made to zero with additional assumptions. Assume $U_{i,t}$ is purely measurement error, then $\mathbf{b}_{1,1,i} = \mathbf{b}_{2,12,ij} = \mathbf{0}$. Sometimes it is assumed that all firms survive from t_1 to t_2 (i.e. there is no sample attrition), or the analyst focuses on a subsample, then $\chi_{i,t_2} = P_{i,t_1} = 1$, hence $\mathbf{b}_{1,2,i} = \mathbf{b}_{2,22,ij} = \mathbf{0}$. ⌟

SA-5.8 Conditional Moment Restrictions

The 2SLS estimator is closely related to another class of problems, those defined by conditional moment restrictions. In this section we consider the following problem:

$$\mathbb{E}[e(Y_i, X_i, \theta_0)|\mathbf{Z}_i] = 0,$$

where Y_i is the outcome variable, X_i is the endogenous regressor, and $\mathbf{Z}_i$ are excluded instruments (or transformations thereof). We can consider more general problems where some covariates are

also included in the estimating equation, which would not change the general conclusion as long as the dimension of these covariates remains fixed.

To transform the above estimating equation into an unconditional form, one can essentially use any function $g(\mathbf{Z}_i)$: $\mathbb{E}[g(\mathbf{Z}_i)e(Y_i, X_i, \theta_0)] = 0$, provided the parameter remains identified. One particular choice is the following:

$$\mathbb{E}[\mu_i e(Y_i, X_i, \theta_0)] = 0, \qquad \mu_i = \mathbf{Z}_i^\mathsf{T}\boldsymbol{\beta}, \tag{E.27}$$

and $\boldsymbol{\beta}$ is the (population) regression coefficient of X_i on $\mathbf{Z}_i$. Note that this reduces to the 2SLS estimator if $e(Y_i, X_i, \theta_0) = Y_i - X_i\theta_0$. 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.27) as given, and investigate how the first-step estimate affects the asymptotic distribution of $\hat{\theta}$, which is given by $\sum_i \hat{\mu}_i e(Y_i, X_i, \hat{\theta}) = 0$. Once again, 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:

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

which has the order $O_\mathbb{P}(k/\sqrt{n})$. The same arguments made in the previous subsection holds here: the bias is essentially a leave-in bias, hence a simple JIVE is effective for bias correction.

The choice of instrument in (E.27) is arbitrary and is not optimal. A more interesting behavior arises when the optimal instrument, under possible conditional heteroskedasticity, is used. The optimal instrument is given by

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

which requires estimating two unknown functions, μ_{1i} and μ_{2i} (see Section SA-4.1 for this generalization). Note that our results apply directly 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. First consider 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.

Now consider 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 nonlinear in μ_{2i}, hence the leading bias is no longer a leave-in bias, and the JIVE is not effective. On the other hand, our generic fully data-driven results do apply, and our proposed jackknife bias-correction and bootstrap-based inference can be used directly.

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

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

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

Step 2. The jackknife bias estimator is defined as

$$\hat{\boldsymbol{\mathcal{B}}} = (n-1) \cdot \sqrt{n} \left(\hat{\boldsymbol{\theta}}^{(\cdot)} - \hat{\boldsymbol{\theta}} \right) = \frac{n-1}{n} \sum_j \sqrt{n} \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}} \right), \tag{E.28}$$

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

$$\hat{\boldsymbol{\mathcal{V}}} = (n-1) \sum_j \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)} \right) \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)} \right)^{\mathsf{T}}. \tag{E.29}$$

⌟

Remark (Notation)**.** To match the notation used in the main paper, note that $\hat{\mathscr{B}} = \hat{\boldsymbol{\mathcal{B}}}/\sqrt{n}$, 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 that $k \propto \sqrt{n}$, hence facilitates the state and proof of relevant results. ⌟

In addition to being fully automatic, another appealing feature of the jackknife is that it is possible to exploit the specific structure of the problem to reduce computation burden. For example, we use a linear model to approximate the unobserved quantity μ_i, the leave-j-out estimate $\hat{\mu}_i^{(j)}$ can easily be obtained by

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

Since re-estimating μ_i is the most time-consuming step when k is large, the above greatly simplifies the algorithm and reduces the computing time. To be more specific, obtaining $\hat{\mu}_i^{(j)}$ naively in each

jackknife repetition requires constructing $(n-1) \times (n-1)$ projection matrix n times, which by using the above method, one only has to construct the projection matrix once.

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

Both A.3(1) and A.3(2) are understood as "design balance", which states that asymptotically the projection matrix is not "concentrated" on any observation. Both are crucial since otherwise $\mathbf{Z}^\top \mathbf{Z}$ becomes singular after deleting that observation. It is weaker than $\max_{1 \le i \le n} \pi_{ii} = o_{\mathbb{P}}(1)$, which is assumed in Mammen (1989) and subsequent work in the area of high-dimensional statistics. In Section SA-2.4, we provide an example in which Assumptions A.3(1) and A.3(2) hold, but $\max_{1 \le i \le n} \pi_{ii}$ has a nondegenerate limiting distribution.

Now we are ready to state the main theorem of this section, concerning validity of the jackknife.

Proposition SA.17 (Jackknife Validity)**.**
Assume A.1, A.2 and A.3 hold, and $k = O(\sqrt{n})$. Then the jackknife bias correction estimate (E.28) and variance estimate (E.29) 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,

$$\begin{aligned}\hat{\mu}_i^\star &= \mathbf{z}_i^\mathsf{T}\left(\sum_j \mathbf{z}_j\mathbf{z}_j^\mathsf{T}\right)^-\left(\sum_j \mathbf{z}_j(\hat{\mu}_j+\hat{\varepsilon}_j\cdot e_j^\star)\right)\\ &= \hat{\mu}_i + \mathbf{z}_i^\mathsf{T}\left(\sum_j \mathbf{z}_j\mathbf{z}_j^\mathsf{T}\right)^-\left(\sum_j \mathbf{z}_j\hat{\varepsilon}_j\cdot e_j^\star\right), \qquad \hat{\varepsilon}_j = r_j-\hat{\mu}_j. \end{aligned} \tag{E.30}$$

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

$$\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})\right]^\mathsf{T}\boldsymbol{\Omega}_n\left[\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i,\hat{\mu}_i^\star,\hat{\boldsymbol{\theta}}^\star)\cdot(1+e_i^\star)\right] = o_\mathbb{P}(1), \tag{E.31}$$

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

Remark (The Multiplier Bootstrap)**.** In principal, one can also use the multiplier bootstrap for the first step, which gives a more unified treatment, with the first step bootstrap as

$$\mathbf{0} = \sum_i (1+e_i^\star)\cdot\mathbf{z}_i(r_i-\mathbf{z}_i^\mathsf{T}\hat{\boldsymbol{\beta}}^\star).$$

To state the above into a two-step procedure, note that one only needs to change the definition of $\hat{\mu}_i^\star$ in (E.31): $\hat{\mu}_i^\star = \mathbf{z}_i^\mathsf{T}(\mathbf{Z}^\mathsf{T}\mathbf{E}^\star\mathbf{Z})^-\mathbf{Z}^\mathsf{T}\mathbf{E}^\star\mathbf{R}$, where $\mathbf{E}^\star$ is a diagonal matrix with diagonal elements

$\{1 + e_i^\star\}_{1 \le i \le n}$. There is one drawback, however, of using the multiplier bootstrap for the first step: for each bootstrap repetition, one has to re-compute the QR decomposition of $\mathbf{E}^{\star\frac{1}{2}}\mathbf{Z}$, which is computationally intensive when k is large. On the other hand, with the first step re-estimated by the wild bootstrap, one only needs to compute the projection matrix $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-}\mathbf{Z}^\mathsf{T}$ once, hence requiring much less computation. ⌟

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 $1 \le \ell \le d_\mu$. 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.30) and (E.31), and is tight.

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

A.4(3) For some $0 < \alpha,\ \delta < \infty$, $\mathcal{H}^{\alpha,\delta}(\partial\dot{\mathbf{m}}_i/\partial\boldsymbol{\theta}) \in \mathsf{BM}_1$. ⌟

Remark (On Assumption A.4(2))**.** Here we give some primitive conditions that imply Assumption A.4(2).

Lemma SA.18 (Primitive Conditions for Assumption A.4(2))**.**
Assume A.1 and A.2 hold. Further assume (i) $\max_{1\le i\le n}\pi_{ii} = O_\mathbb{P}(1/\log(n))$, and (ii) $e_i^\star$ is bounded and has symmetric distribution. Then $\max_{1\le i\le n}|\hat{\mu}_i^\star - \mu_i| = o_\mathbb{P}(1)$. ⌟

To see the intuition, note that it suffices to prove $\max_{1\le i\le n}|\hat{\mu}_i^\star - \hat{\mu}_i| = o_\mathbb{P}(1)$, which is equivalent to $\max_{1\le i\le n}|\sum_j \pi_{ij}e_j^\star\hat{\varepsilon}_j| = o_\mathbb{P}(1)$. Then we make the decomposition:

$$\max_{1\le i\le n}|\sum_j \pi_{ij}e_j^\star\hat{\varepsilon}_j| \le \max_{1\le i\le n}|\sum_j \pi_{ij}e_j^\star\varepsilon_j| + \max_{1\le i\le n}|\sum_j \pi_{ij}e_j^\star(\hat{\mu}_j - \mu_j)|.$$

The second term can be easily handled by Assumption A.1(3) and the condition on $\max_{1\le i\le n}\pi_{ii}$, while for the first term, one needs a reversed symmetrization of (van der Vaart and Wellner, 1996, Lemma 2.3.7). We leave the technical details to the appendix. ⌟

The consistency of $\hat{\boldsymbol{\theta}}^\star$ is very easy to establish:

Proposition SA.19 (Consistency: Bootstrap)**.**
Assume Assumptions A.1(1)–A.1(4) and A.4(1)–A.4(2) hold. Then $|\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}}| = o_{\mathbb{P}}(1)$.

Given consistency, we are able to linearize the estimating equation (E.31) 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:

$$\frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i, \hat{\mu}_i^\star, \hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) \tag{E.32}$$

$$+ \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big) \tag{E.33}$$

$$+ \frac{1}{\sqrt{n}}\sum_i \frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)^2. \tag{E.34}$$

Analyzing the above terms are similar to Lemma SA.3, SA.4, SA.6 and SA.7, with slightly more delicate arguments. We first consider (E.32):

Lemma SA.20 (Term (E.32))**.**
Assume A.1, A.2 and A.4 hold, and $k = O(\sqrt{n})$. Then

$$(\text{E.32}) = \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathbb{P}}(1).$$

⌟

As expected, (E.32) resembles (E.7) and contributes to the variability of $\hat{\boldsymbol{\theta}}^\star$.

For (E.33), we will show that it contributes both to the asymptotic variance as well as the asymptotic bias. Hence it resembles (E.8).

Lemma SA.21 (Term (E.33))**.**
Assume A.1, A.2 and A.4 hold, and $k = O(\sqrt{n})$. Then

$$(\text{E.33}) = \frac{1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)|\mathbf{z}_j\right]\pi_{ij}\right)\varepsilon_i e_i^\star + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_{1,i}$ is given in Lemma SA.4. ⌟

Finally we give the result for (E.34), whose behavior resembles that of (E.9).

Lemma SA.22 (Term (E.34)).
Assume A.1, A.2 and A.4 hold, and $k = O(\sqrt{n})$. Then

$$(\text{E.34}) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + \frac{1}{\sqrt{n}} \sum_{i} \mathbf{b}_{2,ii} \cdot \pi_{ii}^2 \cdot \mathbb{E}[e_i^{\star 3}] + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_{2,ij}$ is given in Lemma SA.7. ⌟

Now we state a result that is similar to Proposition SA.8, by combining Lemma SA.20–SA.22.

Proposition SA.23 (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.8, and

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

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

⌟

Finally we note that without bias, bootstrap consistency can be established easily by appealing to Lindeberg-type CLT arguments, by conditional on the original data. On the other hand, the bootstrap is able to replicate the many covariates/instruments bias only under the assumption that $\mathcal{B}' = \mathbf{0}$, 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 $\mathscr{T} = \left(\hat{\boldsymbol{\mathcal{V}}}/n\right)^{-1/2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \hat{\boldsymbol{\mathcal{B}}}/\sqrt{n}\right)$.

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

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

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

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

Remark (Notation)**.** In the main paper, we use different scaling:

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

hence equivalently, $\mathscr{T}^\star = \left(\hat{\mathscr{V}}^\star\right)^{-1/2}\left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \hat{\mathscr{B}}^\star\right)$. ⌟

Remark (Centering the bootstrap distribution)**.** Asymptotically the distribution of $\mathscr{T}^\star$ is centered at 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 $\mathscr{T}^\star = \left(\hat{\boldsymbol{\mathcal{V}}}^\star/n\right)^{-1/2}\Big(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n} - \mathbb{E}^\star[\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n}]\Big)$. ⌟

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, $\hat{\boldsymbol{\theta}}^{\star,(\ell)} = \arg\min_{\boldsymbol{\theta}} \left|\boldsymbol{\Omega}_n^{1/2}\sum_{i=1, i\neq\ell}^n \omega_i^\star \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{\star,(\ell)}, \boldsymbol{\theta})\right|$. 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.17 and SA.23.

Proposition SA.24 (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. The first data generating process consists of a low dimensional and correctly specified propensity score, while we add redundant covariates to the first step and see the consequence. In the second data generating process, 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. For the third data generating process, 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 sue 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 $\mathbf{Z}_i = [1, \{Z_{\ell,i}\}_{1\leq\ell\leq 199}]$, 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: $T_i = \mathbb{1}\left[0.1 + \sum_{\ell=1}^{4} Z_{\ell,i} \geq V_i\right]$. 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 $1 \leq \ell \leq k - 1$, 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 noises 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 $\tau_{\mathtt{MTE}}(a) = 1 - a$.

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 4th-order interactions
21	$\mathbf{s}^{11}(\mathbf{Z}_i)$ and 2nd-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 5th-order interactions
56	$\mathbf{s}^{26}(\mathbf{Z}_i)$ and 3rd-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, Heckman, and Vytlacil (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.

Outcome Equation

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

(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.

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:

(i) as described above;

(ii) as described above;

(iii) as described above;

(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.

The next specification of $\mathbf{Z}_i$ include certain linear interactions, and corresponds to columns (2), (7) and (12) in Table 11–14:

(i) as described above;

(ii) as described above;

(iii) as described above;

(v) as described above;

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

Another specification of $\mathbf{Z}_i$, corresponding to columns (3), (8) and (13) in Table 11–14, is as follows:

(i) as described above;

(ii) as described above;

(iii) as described above;

(v) as described above;

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

The next specification encompasses all above, given in columns (4), (9) and (14) in 11–14:

(i) as described above;

(ii) as described above;

(iii) as described above;

(v) as described above;

(vi) as described above;

(vii) as described above.

Finally, for comparison purpose, we also include the specification used in Carneiro, Heckman, and Vytlacil (2011), corresponding to columns (4), (10) and (15) in Table 11–14, where the propensity score is estimated with the following Logit regression:

(i) as described above;

(ii) as described above;

(iii) as described above;

(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.

SA-9 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 $80-85\%$.

SA-10 Proofs

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

SA-10.1 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 note that $\pi_{ij} = \pi_{ji}$, 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.25.
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

$$0 \le \pi_{ii} \le 1 \ \forall i, \qquad \text{and} \qquad \sum_i \pi_{ii} = \mathrm{tr}[\mathbf{\Pi}] \le k. \tag{E.35}$$

(iii) The following higher order summations hold:

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

$$\sum_i \pi_{ii}^3 \le \sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j} \pi_{ii}\pi_{ij}^2 \le \sum_i \pi_{ii}^2 \le k, \tag{E.37}$$

$$\sum_i \pi_{ii}^4 \le \sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j} \pi_{ij}^4 \le \sum_i \pi_{ii}^2 \le k, \qquad \sum_{i,j,\ell} \pi_{ij}^2\pi_{i\ell}^2 \le \sum_i \pi_{ii}^2 \le k. \tag{E.38}$$

■

SA-10.2 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.26.

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

■

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

Lemma SA.27.

$$(i,j)\text{-exchangeable} \qquad \left(\sum_{i,j,i\neq j} a_{ij}\right)^2 = \sum_{\substack{i,j,i',j' \\ \text{distinct}}} a_{ij}a_{i'j'} + 4\sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{ii'} + 2\sum_{\substack{i,j \\ i\neq j}} a_{ij}^2. \tag{E.40}$$

■

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.28.

$$(j,\ell)\text{-exchangeable}\left(\sum_{\substack{i,j,\ell\\ \text{distinct}}} a_ib_{ij\ell}\right)^2$$
$$= \sum_i a_i^2\left[\sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{ij'\ell'} + 4\sum_{\substack{j,\ell,j'\\ \text{distinct}}} b_{ij\ell}b_{ijj'} + 2\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}^2\right] + 4\sum_i a_i^2\left[\sum_{j,j\neq i} b_{iij}^2 + \sum_{\substack{j,\ell\\ j\neq i,\ell\neq i,j\neq\ell}} b_{iij}b_{ii\ell}\right]$$
$$- 4\sum_i a_i^2\left[\sum_{\substack{j,\ell,\ell'\\ j\neq\ell,\ell'\neq i}} b_{ij\ell}b_{ii\ell'}\right] + \sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j'\ell'}\right] + 4\sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell,\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j\ell'}\right]$$
$$+ 2\sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}b_{i'j\ell}\right]. \tag{E.41}$$

■

SA-10.3 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.1(3), 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.2(1), Theorem 2.7.11 of van der Vaart and Wellner (1996) implies

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

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-10.4 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

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

where

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

We need uniform law of large numbers (locally to μ_i and θ_0) to show that both $\hat{\mathbf{M}}_n$ and $\tilde{\mathbf{M}}_n$ converge in probability to $\mathbf{M}_0$. Under our assumption, an application of Theorem 2.7.11 of van der Vaart and Wellner (1996) is enough to show that the bracketing number of $\mathbf{m}(\cdot)$ is finite. ■

SA-10.5 Lemma SA.3

The following term is easily bounded:

$$\begin{aligned}
\left|\frac{1}{\sqrt{n}}\sum_i\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\eta_i\right|&\le\max_{1\le i\le n}|\eta_i|\cdot\frac{1}{\sqrt{n}}\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\\
&=o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|=o_{\mathbb{P}}(1),
\end{aligned}$$

since $\dot{\mathbf{m}}$ is integrable. The other term is handled similarly. Note that by Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\Big(\sum_j \pi_{ij}\eta_j\Big)\right| \\
&\le \frac{1}{\sqrt{n}}\left(\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\right)^{1/2}\left(\sum_i\left|\sum_j \pi_{ij}\eta_j\right|^2\right)^{1/2} \\
&\le \frac{1}{\sqrt{n}}\left(\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\right)^{1/2}\left(\sum_i |\eta_i|^2\right)^{1/2} && \text{(projection)} \\
&\le o_{\mathbb{P}}(1)\cdot\left(\frac{1}{n}\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\right)^{1/2} = o_{\mathbb{P}}(1),
\end{aligned}$$

where the last uses the fact that $\dot{\mathbf{m}}$ is square integrable. ■

SA-10.6 Lemma SA.4

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

$$(\text{E.8}) = \frac{1}{\sqrt{n}}\sum_i\left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j,\mu_j,\boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right)\cdot\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij},$$

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.2(4), the above terms are easily bounded by

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

which closes the proof. ■

SA-10.7 Lemma SA.5

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.2(4))

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

where the last line shows why the assumption in Lemma SA.5 is sufficient. Note that by projection, $\boldsymbol{\Gamma}\mathbf{z}_i=\sum_j\boldsymbol{\Gamma}\mathbf{z}_j\pi_{ij}$. ■

SA-10.8 Lemma SA.6

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

$$
\begin{aligned}
\text{(E.9)}&=\frac{1}{2\sqrt{n}}\sum_i\ddot{\mathbf{m}}_i\left(\hat{\mu}_i-\mathbf{z}_i^{\mathsf{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathsf{T}}\boldsymbol{\beta}-\mu_i\right)^2\\
&=\frac{1}{2\sqrt{n}}\sum_i\ddot{\mathbf{m}}_i\left(\hat{\mu}_i-\mathbf{z}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2 \qquad\qquad\text{(I)}
\end{aligned}
$$

$$+ \frac{1}{2\sqrt{n}} \sum_i \ddot{\mathbf{m}}_i \Big(\mathbf{z}_i^\top \boldsymbol{\beta} - \mu_i \Big)^2 \tag{II}$$

$$+ \frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_i \Big(\hat{\mu}_i - \mathbf{z}_i^\top \boldsymbol{\beta} \Big) \Big(\mathbf{z}_i^\top \boldsymbol{\beta} - \mu_i \Big). \tag{III}$$

It is easy to show that (II) is $o_{\mathbb{P}}(1)$ given Assumption A.1(5), A.2(4):

$$|(\mathrm{II})| \le \sqrt{n} \max_{1 \le i \le n} |\eta_i|^2 \cdot \frac{1}{2n} \sum_i |\ddot{\mathbf{m}}_i| = o_{\mathbb{P}}(1).$$

Similar argument applies to (III):

$$\begin{aligned}
|(\mathrm{III})| &= \left| \frac{1}{\sqrt{n}} \sum_{i,j} \ddot{\mathbf{m}}_i \eta_i (\varepsilon_j + \eta_j) \pi_{ij} \right| \\
&\le \frac{1}{\sqrt{n}} \left(\sum_i |\ddot{\mathbf{m}}_i \eta_i|^2 \right)^{1/2} \left(\sum_i |\varepsilon_j + \eta_j|^2 \right)^{1/2} && \text{(projection)} \\
&\le \sqrt{n} \max_{1 \le i \le n} |\eta_i| \cdot \left(\frac{1}{n} \sum_i |\ddot{\mathbf{m}}_i|^2 \right)^{1/2} \left(\frac{1}{n} \sum_i |\varepsilon_j + \eta_j|^2 \right)^{1/2} = o_{\mathbb{P}}(1).
\end{aligned}$$

Next we further decompose (I) as

$$\begin{aligned}
&\frac{1}{2\sqrt{n}} \sum_i \ddot{\mathbf{m}}_i \Big(\hat{\mu}_i - \mathbf{z}_i^\top \boldsymbol{\beta} \Big)^2 \\
&= \frac{1}{2\sqrt{n}} \sum_i \ddot{\mathbf{m}}_i \Big(\sum_j \pi_{ij} \varepsilon_j \Big)^2 && \text{(IV)} \\
&+ \frac{1}{2\sqrt{n}} \sum_i \ddot{\mathbf{m}}_i \Big(\sum_j \pi_{ij} \eta_j \Big)^2 && \text{(V)} \\
&+ \frac{1}{\sqrt{n}} \sum_i \ddot{\mathbf{m}}_i \Big(\sum_{j,k} \varepsilon_j \eta_k \pi_{ij} \pi_{ik} \Big). && \text{(VI)}
\end{aligned}$$

Again we claim that (V) and (VI) are negligible. For (V),

$$\begin{aligned}
|(\mathrm{V})| &\le \left(\max_{1 \le i \le n} \left| \sum_j \pi_{ij} \eta_j \right| \right) \cdot \frac{1}{2\sqrt{n}} \sum_i \left| \ddot{\mathbf{m}}_i \sum_j \pi_{ij} \eta_j \right| \\
&\le \frac{1}{2\sqrt{n}} \left(\max_{1 \le i \le n} \left| \sum_j \pi_{ij} \eta_j \right| \right) \left(\sum_i |\ddot{\mathbf{m}}_i|^2 \right)^{1/2} \left(\sum_i \left(\sum_j \pi_{ij} \eta_j \right)^2 \right)^{1/2} \\
&\le \frac{1}{2\sqrt{n}} \left(\max_{1 \le i \le n} \left| \sum_j \pi_{ij} \eta_j \right| \right) \left(\sum_i |\ddot{\mathbf{m}}_i|^2 \right)^{1/2} \left(\sum_i \eta_i^2 \right)^{1/2} && \text{(projection)} \\
&\precsim \frac{1}{\sqrt{n}} \left(\sqrt{n} \max_{1 \le i \le n} |\eta_i| \right) \cdot O_{\mathbb{P}}(\sqrt{n}) \cdot o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).
\end{aligned}$$

For (VI), note that

$$\begin{aligned}
|(\mathrm{VI})| &\le \left(\frac{1}{n} \sum_i \left| \ddot{\mathbf{m}}_i \Big(\sum_j \varepsilon_j \pi_{ij} \Big) \right|^2 \right)^{1/2} \left(\sum_i \Big(\sum_j \eta_j \pi_{ij} \Big)^2 \right)^{1/2} \\
&\le \left(\frac{1}{n} \sum_i \left| \ddot{\mathbf{m}}_i \Big(\sum_j \varepsilon_j \pi_{ij} \Big) \right|^2 \right)^{1/2} \left(\sum_i |\eta_i|^2 \right)^{1/2} && \text{(projection)} \\
&\le o_{\mathbb{P}}(1) \left(\frac{1}{n} \sum_i \left| \ddot{\mathbf{m}}_i \Big(\sum_j \varepsilon_j \pi_{ij} \Big) \right|^2 \right)^{1/2}.
\end{aligned}$$

Finally, the term in the brackets can be handled with conditional expectation calculation:

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{n}\sum_i\left|\ddot{\mathbf{m}}_i\Big(\sum_j\varepsilon_j\pi_{ij}\Big)\right|^2\right]=\frac{1}{n}\sum_{i,j}\mathbb{E}_{[\cdot|\mathbf{Z}]}[|\ddot{\mathbf{m}}_i\varepsilon_j|^2]\pi_{ij}^2\precsim\frac{k}{n}.$$

Therefore,

$$\text{(VI)}=o_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right).$$

■

SA-10.9 Lemma SA.7

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

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

SA-10.9.1 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.36). And

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

SA-10.9.2 Variance, Term (I)

First for (I) we use (E.41) 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

$$\begin{aligned}&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_{\substack{i,j,\ell\\\text{distinct}}}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell\right)^2\right]\\&=\frac{2}{n}\sum_i\sum_{j,\ell,j\neq\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top\ddot{\mathbf{m}}_ib_{ij\ell}^2\right] &\text{(I.1)}\\&+\frac{4}{n}\sum_i\sum_{j,j\neq i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top\ddot{\mathbf{m}}_ib_{iij}^2\right] &\text{(I.2)}\\&+\frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j,\ell,j\neq\ell}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top\ddot{\mathbf{m}}_{i'}b_{ij\ell}b_{i'jl}\right]. &\text{(I.3)}\end{aligned}$$

Next by (E.38),

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

And by (E.37),

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

And

$$\begin{aligned}
\text{(I.3)} &= \frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j,j',j\neq j'} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_{i'}\pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\varepsilon_j^2\varepsilon_{j'}^2\right] \\
&= \frac{2}{n}\sum_{\substack{i,i',j,j'\\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_{i'}\pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\varepsilon_j^2\varepsilon_{j'}^2\right] \\
&+ \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_{i'}\pi_{ii}\pi_{ii'}^2\pi_{i'i'}\varepsilon_i^2\varepsilon_{i'}^2\right] + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top \ddot{\mathbf{m}}_{i'}\pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\varepsilon_i^2\varepsilon_j^2\right].
\end{aligned}$$

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.41) the above becomes

$$\begin{aligned}
\text{(I.3)} &= \frac{2}{n}\sum_{\substack{i,i',j,j'\\ \text{distinct}}} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\top\mathbf{c}_{i'}d_jd_{j'} + \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\mathbf{e}_i^\top\mathbf{e}_{i'} + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\mathbf{e}_i^\top\mathbf{c}_{i'}d_j \\
&= \frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j,j',j\neq j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\top\mathbf{c}_{i'}d_jd_{j'} \\
&+ \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^\top\mathbf{e}_{i'} - \mathbf{c}_i^\top d_i\mathbf{c}_{i'}d_{i'}\right) + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_id_i\right)^\top\mathbf{c}_{i'}d_j \\
&= \frac{2}{n}\sum_{i,i',j,j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\top\mathbf{c}_{i'}d_jd_{j'} && \text{(I.3.1)} \\
&\quad + \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^\top\mathbf{e}_{i'} - \mathbf{c}_i^\top d_i\mathbf{c}_{i'}d_{i'}\right) && \text{(I.3.2)} \\
&\quad + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_id_i\right)^\top\mathbf{c}_{i'}d_j && \text{(I.3.3)} \\
&\quad - \frac{2}{n}\sum_{i,i',i\neq i'}\sum_j \pi_{ij}^2\pi_{i'j}^2\mathbf{c}_i^\top\mathbf{c}_{i'}d_j^2 && \text{(I.3.4)} \\
&\quad - \frac{2}{n}\sum_i\sum_{j,j'} \pi_{ij}^2\pi_{ij'}^2|\mathbf{c}_i|^2d_jd_{j'}. && \text{(I.3.5)}
\end{aligned}$$

Then use (E.36)

$$\begin{aligned}
|\text{(I.3.1)}| &= \left|\frac{2}{n}\sum_{i,i',j,j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^\top\mathbf{c}_{i'}d_jd_{j'}\right| = \left|\frac{2}{n}\sum_{i,i'}\mathbf{c}_i^\top\mathbf{c}_{i'}\left(\sum_j \pi_{ij}\pi_{i'j}d_j\right)^2\right| \le \max_{1\le i,i'\le n}|\mathbf{c}_i^\top\mathbf{c}_{i'}|\frac{2}{n}\sum_{i,i'}\left(\sum_j \pi_{ij}\pi_{i'j}d_j\right)^2 \\
&= \max_{1\le i,i'\le n}|\mathbf{c}_i^\top\mathbf{c}_{i'}|\frac{2}{n}\sum_{i,i',j,j'} \pi_{ij}\pi_{ij'}d_jd_{j'} = \max_{1\le i,i'\le n}|\mathbf{c}_i^\top\mathbf{c}_{i'}|\frac{2}{n}\sum_{j,j'} d_jd_{j'}\left(\sum_i \pi_{ij}\pi_{ij'}\right)^2 \\
&\le \max_{1\le i,i',j,j'\le n}|\mathbf{c}_i^\top\mathbf{c}_{i'}d_jd_{j'}|\frac{2}{n}\sum_{j,j'}\left(\sum_i \pi_{ij}\pi_{ij'}\right)^2 \le \max_{1\le i,i',j,j'\le n}|\mathbf{c}_i^\top\mathbf{c}_{i'}d_jd_{j'}|\frac{2}{n}\sum_{j,j'}\pi_{jj'}^2 \precsim \frac{k}{n}.
\end{aligned}$$

And by (E.37)

$$|(\text{I.3.2})| = \left| \frac{4}{n} \sum_{\substack{i,i' \\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'} \left(\mathbf{e}_i^\top \mathbf{e}_{i'} - \mathbf{c}_i^\top d_i \mathbf{c}_{i'} d_{i'} \right) \right| \leq \max_{1\leq i,i'\leq n} |\mathbf{e}_i^\top \mathbf{e}_{i'} - \mathbf{c}_i^\top d_i \mathbf{c}_{i'} d_{i'}| \frac{4}{n} \sum_{\substack{i,i' \\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}$$
$$\leq \max_{1\leq i,i'\leq n} |\mathbf{e}_i^\top \mathbf{e}_{i'} - \mathbf{c}_i^\top d_i \mathbf{c}_{i'} d_{i'}| \frac{4}{n} \sum_{\substack{i,i' \\ \text{distinct}}} \pi_{ii'}^2\pi_{i'i'} \precsim \frac{k}{n}.$$

And by (E.36) and (E.38)

$$|(\text{I.3.3})| = \left| \frac{8}{n} \sum_{\substack{i,i',j \\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j} \left(\mathbf{e}_i - \mathbf{c}_i d_i\right)^\top \mathbf{c}_{i'} d_j \right| \precsim \frac{1}{n} \sum_{i',j,i'\neq j} |\mathbf{c}_{i'} d_j||\pi_{i'j}| \left| \sum_{\substack{i \\ i\neq i', i\neq j}} (\mathbf{e}_i - \mathbf{c}_i d_i)\pi_{ii}\pi_{ij}\pi_{ii'} \right|$$
$$\precsim \frac{1}{n} \sqrt{\sum_{i',j} \pi_{i'j}^2} \sqrt{\sum_{i',j} \left| \sum_{\substack{i \\ i\neq i', i\neq j}} (\mathbf{e}_i - \mathbf{c}_i d_i)\pi_{ii}\pi_{ij}\pi_{ii'} \right|^2}$$
$$\precsim \frac{\sqrt{k}}{n} \sqrt{\sum_{i,i',j,j'} (\mathbf{e}_i - \mathbf{c}_i d_i)^\top (\mathbf{e}_{j'} - \mathbf{c}_{j'} d_{j'}) \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{j'j'}\pi_{jj'}\pi_{i'j'}}$$
$$= \frac{\sqrt{k}}{n} \sqrt{\sum_{i,j'} (\mathbf{e}_i - \mathbf{c}_i d_i)^\top (\mathbf{e}_{j'} - \mathbf{c}_{j'} d_{j'}) \pi_{ii}\pi_{ij'}^2\pi_{j'j'}} \precsim \frac{\sqrt{k}}{n} \sqrt{\sum_{i,j'} \pi_{ii}\pi_{ij'}^2\pi_{j'j'}} \leq \frac{\sqrt{k}}{n} \sqrt{\sum_{i,j'} \pi_{ii}\pi_{ij'}^2} \leq \frac{\sqrt{k}}{n} \sqrt{\sum_i \pi_{ii}^2} \leq \frac{k}{n}.$$

And by (E.38)

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

And by (E.38)

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

SA-10.9.3 Variance, Term (II)

Then for (II), one has (by using (E.39))

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

$$+\frac{1}{n}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[|\ddot{\mathbf{m}}_i|^2\pi_{ij}^4\varepsilon_j^4\right] \tag{II.5}$$

$$+\frac{1}{n}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top\ddot{\mathbf{m}}_j\pi_{ij}^4\varepsilon_i^2\varepsilon_j^2\right]. \tag{II.6}$$

With (E.38) 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 $O_{\mathbb{P}}(n^{-1}\sum_i\pi_{ii}^2)=O_{\mathbb{P}}(k/n)$.

SA-10.9.4 Variance, Term (III)

Next we consider (III), and still (E.39) implies

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{ii}\varepsilon_j\varepsilon_i\right)^2\right]\\
&=\frac{4}{n}\sum_{i,j,\text{distinct}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[|\ddot{\mathbf{m}}_i|^2\pi_{ij}^2\pi_{ii}^2\varepsilon_i^2\varepsilon_j^2\right] &\text{(III.1)}\\
&+\frac{8}{n}\sum_{i,j,\text{distinct}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^\top\ddot{\mathbf{m}}_j\pi_{ij}^2\pi_{ii}\pi_{jj}\varepsilon_i^2\varepsilon_j^2\right]. &\text{(III.2)}
\end{aligned}$$

For (III.1) it is bounded by

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

which is bounded by k/n due to (E.37). 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.37) and $\pi_{ii}\leq 1$.

SA-10.9.5 Variance, Term (IV)

Finally we consider (IV), and the variance is

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

And both (IV.1) and (IV.2) 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.9). This is trivial due to Assumption A.2(4), and the consistency assumption A.1(3). ■

SA-10.10 Theorem SA.8

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.8) and (E.9) are of order $O_{\mathbb{P}}(1)$. By Assumption A.1(3), the remainder term in the quadratic expansion (after (E.9)) has the order $o_{\mathbb{P}}(|(\text{E.9})|)$, which is negligible. ■

SA-10.11 Theorem SA.9

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

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

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:

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

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:

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

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) $\tilde{\mathbf{\Psi}}_1$ converges unconditionally to a multivariate normal distribution; and (2) conditional on $\mathbf{Z}$, $\tilde{\mathbf{\Psi}}_2$ converges to a multivariate normal distribution (more precisely, conditional on $\mathbf{Z}$ the distribution function of $\tilde{\mathbf{\Psi}}_2$ 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

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

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

Hence we are able to show

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

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-10.12 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 $\hat{\boldsymbol{\theta}}$. Then we have

$$\sqrt{n}\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big)=\Big(\mathbb{E}\mathbb{V}[\mathbf{f}_i|\mathbf{z}_i]\Big)^{-1}\left[\frac{1}{\sqrt{n}}\sum_i\Big(\sum_j q_{ij}\mathbf{f}(\mathbf{x}_j,\hat{\mu}_j,\boldsymbol{\theta}_0)\Big)\Big(y_i-f(\mathbf{x}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\Big)\right]+o_{\mathbb{P}}(1),$$

which we now analyze. Again we linearize with respect to the first step estimate to the second order:

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

Term (I) has the following further expansion:

$$\text{(I)}=\frac{1}{\sqrt{n}}\sum_{i,j}(\mathbf{f}_i-\pi_{ij}\mathbf{f}_j)u_i=\frac{1}{\sqrt{n}}\sum_i\mathbf{f}_iu_i-\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{f}_ju_i\pi_{ij}$$

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

The last line uses two facts: first $\mathbf{z}_i$ can approximate the expectation $\mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]$, second u_i has zero conditional expectation, so that there is no bias contribution from (I).

For (II), we consider the following:

$$\text{(II)} + \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{f}}_i u_i \Big(\hat{\mu}_i - \mu_i\Big) \tag{II.1}$$
$$- \frac{1}{\sqrt{n}} \sum_{i,j} \dot{\mathbf{f}}_i u_j \pi_{ij} \Big(\hat{\mu}_i - \mu_i\Big) \tag{II.2}$$
$$- \frac{1}{\sqrt{n}} \sum_i \mathbf{f}_i \dot{f}_i \Big(\hat{\mu}_i - \mu_i\Big) \tag{II.3}$$
$$+ \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{f}_j \dot{f}_i \pi_{ij} \Big(\hat{\mu}_i - \mu_i\Big). \tag{II.4}$$

Then

$$\text{(II.1)} = \frac{1}{\sqrt{n}} \sum_{ij} \dot{\mathbf{f}}_i u_i \varepsilon_j \pi_{ij} = \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\dot{\mathbf{f}}_i u_i \varepsilon_i|\mathbf{z}_i] \pi_{ii} + o_{\mathbb{P}}(1),$$

which is bias contribution. And

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

so that this term has bias contribution. Next

$$\text{(II.3)} = -\frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{f}_i \dot{f}_i \varepsilon_j \pi_{ij} = -\frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}[\mathbf{f}_i \dot{f}_i|\mathbf{z}_i] \varepsilon_j \pi_{ij} - \frac{1}{\sqrt{n}} \sum_{i,j} (\mathbf{f}_i \dot{f}_i - \mathbb{E}[\mathbf{f}_i \dot{f}_i|\mathbf{z}_i]) \varepsilon_j \pi_{ij}$$
$$= -\frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\mathbf{f}_i \dot{f}_i|\mathbf{z}_i] \varepsilon_i - \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\mathbf{f}_i \dot{f}_i \varepsilon_i|\mathbf{z}_i] \pi_{ii} + o_{\mathbb{P}}(1),$$

hence this term has both variance and bias contribution. Finally

$$\text{(II.4)} = \frac{1}{\sqrt{n}} \sum_{i,j,k} \mathbf{f}_j \dot{f}_i \varepsilon_k \pi_{ij} \pi_{ik} = \frac{1}{\sqrt{n}} \sum_{i,j,k} \mathbb{E}[\mathbf{f}_j|\mathbf{z}_j] \dot{f}_i \varepsilon_k \pi_{ij} \pi_{ik} + \frac{1}{\sqrt{n}} \sum_{i,j,k} (\mathbf{f}_j - \mathbb{E}[\mathbf{f}_j|\mathbf{z}_j]) \dot{f}_i \varepsilon_k \pi_{ij} \pi_{ik}$$
$$= \frac{1}{\sqrt{n}} \sum_{i,k} \mathbb{E}[\mathbf{f}_i|\mathbf{z}_i] \dot{f}_i \varepsilon_k \pi_{ik} + \frac{1}{\sqrt{n}} \sum_{i,j,k} (\mathbf{f}_j - \mathbb{E}[\mathbf{f}_j|\mathbf{z}_j]) \dot{f}_i \varepsilon_k \pi_{ij} \pi_{ik} + o_{\mathbb{P}}(1)$$
$$= \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\mathbf{f}_i|\mathbf{z}_i] \mathbb{E}[\dot{f}_i|\mathbf{z}_i] \varepsilon_i + \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\mathbf{f}_i|\mathbf{z}_i] \mathbb{E}[\dot{f}_i \varepsilon_i|\mathbf{z}_i] \pi_{ii} + \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}[\dot{f}_i \mathbf{f}_j \varepsilon_j|\mathbf{z}_i, \mathbf{z}_j] \pi_{ij}^2 + o_{\mathbb{P}}(1),$$

so that this term has both variance and bias contribution.

Term (III) is again split as

$$\text{(III)} = \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{f}}_i u_i \Big(\hat{\mu}_i - \mu_i\Big)^2 \tag{III.1}$$
$$- \frac{1}{\sqrt{n}} \sum_{i,j} \frac{1}{2} \ddot{\mathbf{f}}_i u_j \pi_{ij} \Big(\hat{\mu}_i - \mu_i\Big)^2 \tag{III.2}$$
$$- \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \mathbf{f}_i \ddot{f}_i \Big(\hat{\mu}_i - \mu_i\Big)^2 \tag{III.3}$$
$$+ \frac{1}{\sqrt{n}} \sum_{i,j} \frac{1}{2} \mathbf{f}_j \ddot{f}_i \pi_{ij} \Big(\hat{\mu}_i - \mu_i\Big)^2. \tag{III.4}$$

Then first,

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

due to Assumption A.3(1), hence there is neither variance nor bias contribution from this term. Next,

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

which is a bias contribution. Next,

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

which is a bias contribution. Finally,

$$\begin{aligned}
\text{(III.4)} &= \frac{1}{\sqrt{n}} \sum_{i,j,k,\ell} \frac{1}{2} \mathbf{f}_j \ddot{f}_i \varepsilon_k \varepsilon_\ell \pi_{ij} \pi_{ik} \pi_{i\ell} \\
&= \frac{1}{\sqrt{n}} \sum_{i,k,\ell} \frac{1}{2} \mathbb{E}[\mathbf{f}_i | \mathbf{z}_i] \ddot{f}_i \varepsilon_k \varepsilon_\ell \pi_{ik} \pi_{i\ell} + \frac{1}{\sqrt{n}} \sum_{i,j,k,\ell} \frac{1}{2} (\mathbf{f}_j - \mathbb{E}[\mathbf{f}_j | \mathbf{z}_j]) \ddot{f}_i \varepsilon_k \varepsilon_\ell \pi_{ij} \pi_{ik} \pi_{i\ell} \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \frac{1}{2} \mathbb{E}[\mathbf{f}_i | \mathbf{z}_i] \mathbb{E}[\ddot{f}_i \varepsilon_j^2 | \mathbf{z}_i, \mathbf{z}_j] \pi_{ij}^2 + \frac{1}{\sqrt{n}} \sum_{i,j} \frac{1}{2} \mathbb{E}[\ddot{f}_i \mathbf{f}_j \varepsilon_j^2 | \mathbf{z}_i, \mathbf{z}_j] \pi_{ij}^3 + o_{\mathbb{P}}(1),
\end{aligned}$$

which is again a bias contribution.

Finally for (IV), it is decomposed as

$$\text{(IV)} = -\frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{f}}_i \dot{f}_i \Big(\hat{\mu}_i - \mu_i\Big)^2 \tag{IV.1}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i,j} \dot{\mathbf{f}}_j \dot{f}_i \Big(\hat{\mu}_i - \mu_i\Big)\Big(\hat{\mu}_j - \mu_j\Big) \pi_{ij}. \tag{IV.2}$$

First,

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

which is a bias contribution. Then

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

Here we make a complementary calculation:

$$\sum_{i,j} |\pi_{ij} \pi_{ii} \pi_{jj}| \leq \left(\sum_{i,j} \pi_{ii} \right)^{1/2} \left(\sum_i \left(\sum_j \pi_{ij} \pi_{jj} \right)^2 \right)^{1/2} \tag{Cauchy-Schwarz}$$

$$= \left(\sum_i \pi_{ii}\right)^{1/2} \left(\sum_{i,j,k} \pi_{ij}\pi_{jj}\pi_{ik}\pi_{kk}\right)^{1/2} = \left(\sum_i \pi_{ii}\right)^{1/2} \left(\sum_{i,j} \pi_{ij}\pi_{ii}\pi_{jj}\right)^{1/2}.$$

Now we collect terms.

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_i \Big(\sum_j q_{ij} \mathbf{f}(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) \Big(y_i - f(\mathbf{x}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big) = o_{\mathbb{P}}(1) \\
&+ \frac{1}{\sqrt{n}} \sum_i (\mathbf{f}_i - \mathbb{E}[\mathbf{f}_i|\mathbf{z}_i]) u_i - \frac{1}{\sqrt{n}} \sum_i \mathbb{C}\mathrm{ov}[\mathbf{f}_i, \dot{f}_i|\mathbf{z}_i]\varepsilon_i \\
&+ \frac{1}{\sqrt{n}} \sum_i \Big(\mathbb{E}[\dot{\mathbf{f}}_i u_i \varepsilon_i|\mathbf{z}_i] - \mathbb{C}\mathrm{ov}[\mathbf{f}_i, \dot{f}_i \varepsilon_i|\mathbf{z}_i] \Big) \pi_{ii} \\
&+ \frac{1}{\sqrt{n}} \sum_{i,j} \Big(\mathbb{E}[\dot{f}_i|\mathbf{z}_i]\mathbb{E}[\mathbf{f}_j\varepsilon_j|\mathbf{z}_j] - \mathbb{E}[\dot{\mathbf{f}}_i|\mathbf{z}_i]\mathbb{E}[u_j\varepsilon_j|\mathbf{z}_j] - \frac{1}{2}\mathbb{C}\mathrm{ov}[\mathbf{f}_i, \ddot{f}_i|\mathbf{z}_i]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_i, \mathbf{z}_j] - \mathbb{E}[\dot{\mathbf{f}}_i \dot{f}_i \varepsilon_j^2|\mathbf{z}_i, \mathbf{z}_j] \Big) \pi_{ij}^2 \\
&+ \frac{1}{\sqrt{n}} \sum_{i,j,k} \mathbb{E}[\dot{f}_i|\mathbf{z}_i]\mathbb{E}[\dot{\mathbf{f}}_k|\mathbf{z}_k]\mathbb{E}[\varepsilon_j^2|\mathbf{z}_j]\pi_{ij}\pi_{ik}\pi_{jk} \\
&+ \frac{1}{\sqrt{n}} \sum_{i,j} \Big(\frac{1}{2}\mathbb{E}[\ddot{f}_i|\mathbf{z}_i]\mathbb{E}[\mathbf{f}_j\varepsilon_j^2|\mathbf{z}_j] - \frac{1}{2}\mathbb{E}[\ddot{\mathbf{f}}_i|\mathbf{z}_i]\mathbb{E}[u_j\varepsilon_j^2|\mathbf{z}_j] + \mathbb{E}[\dot{f}_i\varepsilon_i|\mathbf{z}_i]\mathbb{E}[\dot{\mathbf{f}}_j\varepsilon_j|\mathbf{z}_j] \Big) \pi_{ij}^3.
\end{aligned}$$

SA-10.13 Proposition SA.10

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

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

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

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

Finally we consider the variance contribution, which utilizes

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

■

SA-10.14 Proposition SA.12

$\hat{\theta}$ has the expansion

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

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-10.15 Proposition SA.13

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 $\mu_i = P_i$)

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

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.1(5), 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:

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

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:

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

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-10.16 Proposition SA.14

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,

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

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-10.17 Proposition SA.15

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 $\dot{\mathbf{X}}_i = -\mathbf{e}_2 = [0,-1]^\top$. Hence

$$\begin{aligned}\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) &= \frac{\partial}{\partial\mu_i}\left\{\mathbf{X}_iL'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big)\right\} = \dot{\mathbf{X}}_iL'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big) + \mathbf{X}_i\dot{\mathbf{X}}_i^\top\boldsymbol{\theta}_0L''(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big) - \mathbf{X}_i\dot{\mathbf{X}}_i^\top\boldsymbol{\theta}_0 \\ &= -\mathbf{e}_2L'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big) + \gamma_0\mathbf{X}_iL''(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big) - \gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)^2.\end{aligned}$$

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

$$\mathbf{b}_{1,i} = \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_i|\mathbf{Z}_i\right] = -\mathbb{E}[\gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)^2\varepsilon_i|\mathbf{Z}_i],$$

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

$$\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\mathbf{Z}_i] = -\mathbb{E}[\gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)^2|\mathbf{Z}_i].$$

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

$$\begin{aligned}\mathbf{b}_{2,ij} &= \mathbb{E}\left[\frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right] = \mathbb{E}\left[\frac{1}{2}\frac{\partial}{\partial\mu_i}\left[-\mathbf{e}_2L'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big)\right]\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right] \\ &\quad + \mathbb{E}\left[\frac{1}{2}\frac{\partial}{\partial\mu_i}\left[\gamma_0\mathbf{X}_iL''(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big(Y_i - L(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\Big)\right]\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right] + \mathbb{E}\left[\frac{1}{2}\frac{\partial}{\partial\mu_i}\left[-\gamma_0\mathbf{X}_iL'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)^2\right]\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right] \\ &= \mathbb{E}\left[\frac{\gamma_0}{2}\mathbf{e}_2L'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)^2\varepsilon_j^2 - \frac{\gamma_0^2}{2}\mathbf{X}_iL''(\mathbf{X}_i^\top\boldsymbol{\theta}_0)L'(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\varepsilon_j^2 - \gamma_0^2\mathbf{X}_iL''(\mathbf{X}_i^\top\boldsymbol{\theta}_0)\varepsilon_j^2\middle|\mathbf{Z}_i,\mathbf{Z}_j\right].\end{aligned}$$

■

SA-10.18 Proposition SA.16

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 $\mathbf{z}_{1i}\supset\mathbf{z}_{2i}$),

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

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

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[1].

[1] 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

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, we remark that V_{i,t_2} may not be orthogonal to time-t_2 information, hence it is generally impossible to conclude the last line being 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

$$\begin{aligned}\mathbf{b}_{2,11,ij} &= -\frac{1}{2}\left\{2\begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1} \end{bmatrix} g_{2,i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{22,i,t_1}\right\} \mathbb{V}\Big[U_{i,t_1}\Big|(L, I, K, A)_{i,t_1}\Big], \\ \mathbf{b}_{2,22,ij} &= -\frac{1}{2}\left\{2\begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1} \end{bmatrix} g_{1,i,t_1} + \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{11,i,t_1}\right\} \mathbb{V}\Big[\chi i, t_2\Big|(L, I, K, A)_{i,t_1}\Big], \\ \mathbf{b}_{2,12,ij} &= -\left\{-\begin{bmatrix} K_{i,t_1} g_{22,i,t_1} \\ A_{i,t_1} g_{22,i,t_1} \\ -\mathbf{g}_{23,i,t_1} \end{bmatrix} g_{1,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{12,i,t_1} \\ A_{i,t_1} g_{12,i,t_1} \\ -\mathbf{g}_{13,i,t_1} \end{bmatrix} g_{2,i,t_1} - \begin{bmatrix} K_{i,t_1} g_{2,i,t_1} - K_{i,t_2} \\ A_{i,t_1} g_{2,i,t_1} - A_{i,t_2} \\ -\mathbf{g}_{3,i,t_1} \end{bmatrix} g_{12,i,t_1}\right\} \mathbb{C}\mathrm{ov}\Big[U_{i,t_1}, \chi_{i,t_2}\Big|(L, I, K, A)_{i,t_1}\Big].\end{aligned}$$

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 unit of aging is calendar year.

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, $\bar{\boldsymbol{\Psi}}_2$, can be decomposed into three, and two of them correspond to contributions of estimating ν_{1i} and μ_{2i} in the first step,

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

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

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

$$\bar{\boldsymbol{\Psi}}_{2,3} = \frac{1}{\sqrt{n}} \frac{1}{\mathbb{E}\mathbb{V}[L_{i,t_1}|(I,K,A)_{i,t_1}]} \tilde{\boldsymbol{\Sigma}}_0 \sum_i \mathbb{E}\Big[L_{i,t_1}\Big|(I,K,A)_{i,t_1}\Big] U_{i,t_1}.$$

■

SA-10.19 Proposition SA.17: 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

$$\begin{aligned}
\mathbf{a}_i &= -\left(\mathbf{M}_0^\intercal \boldsymbol{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \boldsymbol{\Omega}_0 \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \\
\mathbf{b}_i &= -\left(\mathbf{M}_0^\intercal \boldsymbol{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \boldsymbol{\Omega}_0 \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \\
\mathbf{c}_i &= -\frac{1}{2}\left(\mathbf{M}_0^\intercal \boldsymbol{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \boldsymbol{\Omega}_0 \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0).
\end{aligned}$$

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

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

hence

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(\cdot)} - \boldsymbol{\theta}_0\right) &= \frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{a}_i \\
&\quad + \frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\left(\hat{\mu}_i^{(j)} - \mu_i\right) + \frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\hat{\mu}_i^{(j)} - \mu_i\right)^2.
\end{aligned}$$

To simplify, note that

$$\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{a}_i = \frac{\sqrt{n}}{n(n-1)}\sum_i\sum_{j,j\neq i}\mathbf{a}_i = \frac{1}{\sqrt{n}}\sum_i \mathbf{a}_i,$$

and

$$
\begin{aligned}
&\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\left(\hat{\mu}_i^{(j)}-\mu_i\right)\\
&=\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\left[(\hat{\mu}_i-\mu_i)+\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-r_j\right)\right]\\
&=\frac{1}{\sqrt{n}}\sum_i\mathbf{b}_i\left(\hat{\mu}_i-\mu_i\right)+\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-r_j\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\hat{\mu}_i^{(j)}-\mu_i\right)^2\\
&=\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left[(\hat{\mu}_i-\mu_i)+\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-r_j\right)\right]^2\\
&=\frac{1}{\sqrt{n}}\sum_i\mathbf{c}_i\left(\hat{\mu}_i-\mu_i\right)^2\\
&\quad+\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-r_j\right)^2\\
&\quad+\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i-\mu_i\right)\left(\hat{\mu}_j-r_j\right).
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(\cdot)}-\boldsymbol{\theta}_0\right)=\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)&\\
&+\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-r_j\right)\\
&+\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-r_j\right)^2\\
&+\frac{\sqrt{n}}{n(n-1)}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i-\mu_i\right)\left(\hat{\mu}_j-r_j\right).
\end{aligned}
$$

Or equivalently,

$$
\begin{aligned}
(n-1)\cdot\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(\cdot)}-\hat{\boldsymbol{\theta}}\right)=\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-r_j\right)&\qquad\text{(I)}\\
+\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-r_j\right)^2&\qquad\text{(II)}\\
+\frac{2}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i-\mu_i\right)\left(\hat{\mu}_j-r_j\right).&\qquad\text{(III)}
\end{aligned}
$$

By Assumption A.1(5), we could ignore the approximation error. And (I) becomes

$$
\begin{aligned}
\text{(I)}&=\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-\mu_j+\mu_j-r_j\right)\\
&=\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\sum_\ell\pi_{j\ell}\varepsilon_\ell\right)&\text{(I.1)}\\
&\quad-\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\varepsilon_j+o_{\mathbb{P}}(1).&\text{(I.2)}
\end{aligned}
$$

Then we have the following conditional expectations:

$$\begin{aligned}
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\mathrm{I}.1)\right] &= \frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\frac{\pi_{ij}^2}{1-\pi_{jj}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i\varepsilon_i\right] \\
&= -\frac{1}{\sqrt{n}}\left(\mathbf{M}_0^\intercal\mathbf{\Omega}_0\mathbf{M}_0\right)^{-1}\mathbf{M}_0^\intercal\mathbf{\Omega}_0\left[\sum_i\mathbf{b}_{1,i}\pi_{ii}\right]+\frac{1}{\sqrt{n}}\sum_i\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i\varepsilon_i\right]\left(\sum_{j,j\neq i}\frac{\pi_{ij}^2}{1-\pi_{jj}}-\pi_{ii}\right) \\
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\mathrm{I}.2)\right] &= 0.
\end{aligned}$$

To further simplify, note that

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_i\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i\varepsilon_i\right]\left(\sum_{j,j\neq i}\frac{\pi_{ij}^2}{1-\pi_{jj}}-\pi_{ii}\right)\right| \precsim \frac{1}{\sqrt{n}}\sum_i\left|\sum_{j,j\neq i}\frac{\pi_{ij}^2}{1-\pi_{jj}}-\pi_{ii}\right| \\
&\leq \frac{1}{\sqrt{n}}\sum_{i,j}\left|\frac{\pi_{ij}^2}{1-\pi_{jj}}-\pi_{ij}^2\right|+\frac{1}{\sqrt{n}}\sum_i\pi_{ii}^2 && (\pi_{ii}=\textstyle\sum_j\pi_{ij}^2) \\
&= \frac{1}{\sqrt{n}}\sum_{i,j}\frac{\pi_{ij}^2\pi_{jj}}{1-\pi_{jj}}+\frac{1}{\sqrt{n}}\sum_i\pi_{ii}^2 \\
&\precsim_{\mathbb{P}} \frac{1}{\sqrt{n}}\sum_{i,j}\pi_{ij}^2\pi_{jj}+\frac{1}{\sqrt{n}}\sum_i\pi_{ii}^2 \\
&\precsim \frac{1}{\sqrt{n}}\sum_i\pi_{ii}^2 = o_{\mathbb{P}}(1). && \text{(E.37)}
\end{aligned}$$

One could conduct variance calculation, which is tedious yet straightforward. Now we consider (II), which has the following expansion:

$$\begin{aligned}
(\mathrm{II}) &= \frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-\mu_j+\mu_j-r_j\right)^2 \\
&= \frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-\mu_j\right)^2 \\
&\quad+\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\mu_j-r_j\right)^2 \\
&\quad+\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-\mu_j\right)(\mu_j-r_j) \\
&= \frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\sum_{\ell,m}\pi_{j\ell}\pi_{jm}\varepsilon_\ell\varepsilon_m\right) && \text{(II.1)} \\
&\quad+\frac{1}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\varepsilon_j^2 && \text{(II.2)} \\
&\quad-\frac{2}{\sqrt{n}}\sum_j\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\sum_\ell\pi_{j\ell}\varepsilon_\ell\varepsilon_j\right)+o_{\mathbb{P}}(1). && \text{(II.3)}
\end{aligned}$$

Therefore

$$\begin{aligned}
\left|\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\mathrm{II}.1)\right]\right| &= \left|\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\sum_\ell\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i\varepsilon_\ell^2\right]\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{j\ell}^2\right| \\
&\precsim \frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\sum_\ell\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{j\ell}^2 \\
&\precsim_{\mathbb{P}} \frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\sum_\ell\pi_{ij}^2\pi_{j\ell}^2
\end{aligned}$$

$$\leq \frac{1}{\sqrt{n}} \sum_{j,\ell} \pi_{j\ell}^2 \pi_{jj} = o_{\mathbb{P}}(1), \tag{E.37}$$

and

$$\begin{aligned}
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\text{II.2})\right] &= \frac{1}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + o_{\mathbb{P}}(1) \\
&= -\frac{1}{\sqrt{n}} \left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \left[\sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2\right] + \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \frac{\pi_{ij}^2 \pi_{jj}}{(1-\pi_{jj})^2} + o_{\mathbb{P}}(1) \\
&= -\frac{1}{\sqrt{n}} \left(\mathbf{M}_0^\intercal \mathbf{\Omega}_0 \mathbf{M}_0\right)^{-1} \mathbf{M}_0^\intercal \mathbf{\Omega}_0 \left[\sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2\right] + o_{\mathbb{P}}(1),
\end{aligned} \tag{E.37}$$

and

$$\begin{aligned}
\left|\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\text{II.3})\right]\right| &= \left| \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \pi_{jj} \right| \\
&\precsim_{\mathbb{P}} \frac{1}{\sqrt{n}} \sum_{i,j} \pi_{ij}^2 \pi_{jj} = o_{\mathbb{P}}(1).
\end{aligned} \tag{E.37}$$

Finally (III) has the expansion:

$$\begin{aligned}
(\text{III}) &= \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\hat{\mu}_i - \mu_i\right) \left(\hat{\mu}_j - \mu_j + \mu_j - r_j\right) \\
&= \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_\ell \pi_{i\ell} \varepsilon_\ell\right) \left(\sum_m \pi_{jm} \varepsilon_m - \varepsilon_j\right) + o_{\mathbb{P}}(1) \\
&= \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_{\ell,m} \pi_{i\ell} \pi_{jm} \varepsilon_\ell \varepsilon_m\right) && \text{(III.1)} \\
&\quad - \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_\ell \pi_{i\ell} \varepsilon_\ell \varepsilon_j\right) + o_{\mathbb{P}}(1). && \text{(III.2)}
\end{aligned}$$

Again we consider the conditional expectations:

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.1}\right] = \frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \sum_\ell \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{ij} \pi_{i\ell} \pi_{j\ell}}{1-\pi_{jj}},$$

and

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.2}\right] = -\frac{2}{\sqrt{n}} \sum_j \sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \frac{\pi_{ij}^2}{1-\pi_{jj}}.$$

Therefore

$$\begin{aligned}
&\left|\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.1}\right] + \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.2}\right]\right| \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \sum_\ell \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{ij} \pi_{i\ell} \pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \frac{\pi_{ij}^2}{1-\pi_{jj}} \right| \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \sum_\ell \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{ij} \pi_{i\ell} \pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,\ell,i\neq \ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{i\ell}^2}{1-\pi_{\ell\ell}} \right|
\end{aligned}$$

$$
\begin{aligned}
&= \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,\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\mathbf{c}_i \varepsilon_\ell^2 \right] \pi_{i\ell}^2 \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) \\
&\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). && ((\text{E.37}) \text{ and } \pi_{j'j'} \leq 1)
\end{aligned}
$$

Therefore we showed the desired result. ■

SA-10.20 Proposition SA.17: 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

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}_0 &= \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{a}_i + \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{b}_i \left(\hat{\mu}_i^{(j)} - \mu_i \right) + \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{c}_i \left(\hat{\mu}_i^{(j)} - \mu_i \right)^2 \\
&= \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{a}_i + \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{b}_i \left(\hat{\mu}_i + \frac{\pi_{ij}}{1-\pi_{jj}} \left(\hat{\mu}_j - r_j \right) - \mu_i \right) \\
&\quad + \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{c}_i \left(\hat{\mu}_i + \frac{\pi_{ij}}{1-\pi_{jj}} \left(\hat{\mu}_j - r_j \right) - \mu_i \right)^2 \\
&= \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{a}_i + \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{b}_i \left(\hat{\mu}_i - \mu_i \right) + \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\hat{\mu}_j - r_j \right) + \frac{1}{n-1} \sum_{i,i\neq j} \mathbf{c}_i \left(\hat{\mu}_i - \mu_i \right)^2
\end{aligned}
$$

$$+\frac{1}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-r_j\right)^2+\frac{2}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i-\mu_i\right)\left(\hat{\mu}_j-r_j\right).$$

Equivalently,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{(j)}-\hat{\boldsymbol{\theta}}=&\frac{1}{n-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) && \text{(I)}\\
&-\frac{1}{n-1}\mathbf{a}_j && \text{(II)}\\
&-\frac{1}{n-1}\mathbf{b}_j\left(\hat{\mu}_j-\mu_j\right) && \text{(III)}\\
&-\frac{1}{n-1}\mathbf{c}_j\left(\hat{\mu}_j-\mu_j\right)^2 && \text{(IV)}\\
&+\frac{1}{n-1}\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-r_j\right) && \text{(V)}\\
&+\frac{1}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-r_j\right)^2 && \text{(VI)}\\
&+\frac{2}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i-\mu_i\right)\left(\hat{\mu}_j-r_j\right). && \text{(VII)}
\end{aligned}$$

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.8 with

$$\begin{aligned}
(n-1)\sum_j(\text{II})^2&=\mathbb{V}[\bar{\boldsymbol{\Psi}}_1]+o_{\mathbb{P}}(1)\\
(n-1)\sum_j(\text{II})(\text{V})^\mathsf{T}&=\mathbb{C}\text{ov}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_1,\bar{\boldsymbol{\Psi}}_2]+o_{\mathbb{P}}(1)\\
(n-1)\sum_j(\text{V})^2&=\mathbb{V}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_2]+o_{\mathbb{P}}(1).
\end{aligned}$$

Furthermore, all the other square terms and interactions are asymptotically negligible. In the following proof, we use two facts repeatedly, which are collected here. First by the uniform consistency assumption, $\hat{\mu}_i-\mu_i=o_{\mathbb{P}}(1)$ uniformly in i. Second for two sequences $\{u_i\}$ and $\{v_j\}$,

$$\left|\sum_{i,j}u_i\pi_{ij}v_j\right|\leq\sqrt{\sum_i u_i^2}\sqrt{\sum_i\left(\sum_j\pi_{ij}v_j\right)^2}\leq\sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2},$$

where the first inequality comes from the Cauchy-Schwarz inequality, and the second inequality comes from the fact that π_{ij} are elements of projection matrix.

Term (I) is the easiest:

$$(n-1)\sum_j(\text{I})^2=\frac{1}{n-1}\sum_j\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^2\asymp\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^2=o_{\mathbb{P}}(1),$$

by consistency. Then it is also easy to show that for $\dagger=\text{II},\cdots,\text{VII}$

$$(n-1)\sum_j(\text{I})(\dagger)^\mathsf{T}=(\text{I})\frac{1}{n-1}\sum_j(\dagger)^\mathsf{T}=o_{\mathbb{P}}(1)\cdot\frac{1}{n-1}\sum_j(\dagger)^\mathsf{T}=o_{\mathbb{P}}(1),$$

since the summands are bounded in probability uniformly in j.

Next we consider (II):

$$(n-1)\sum_j(\text{II})^2=\frac{1}{n-1}\sum_j\mathbf{a}_j^2,$$

which is asymptotically equivalent to $\mathbb{V}[\bar{\mathbf{\Psi}}_1]$ in Theorem SA.8. Now we consider the interactions:

$$\left|(n-1)\sum_j (\mathrm{II})(\mathrm{III})^\mathsf{T}\right| = \left|\frac{1}{n-1}\sum_j \mathbf{a}_j\mathbf{b}_j^\mathsf{T}\left(\hat{\mu}_j-\mu_j\right)\right| \le o_\mathbb{P}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{a}_j\mathbf{b}_j^\mathsf{T}| = o_\mathbb{P}(1).$$

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):

$$\begin{aligned}
(n-1)\sum_j (\mathrm{II})(\mathrm{V})^\mathsf{T} &= -\frac{1}{n-1}\sum_j \mathbf{a}_j\left(\hat{\mu}_j-r_j\right)\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}} \\
&= -\frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}} + o_\mathbb{P}(1) \qquad \text{(Assumption A.1(3))} \\
&= \frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\pi_{ij} - \frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} + o_\mathbb{P}(1) \\
&= \frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\pi_{ij} + o_\mathbb{P}(1),
\end{aligned}$$

which is asymptotically equivalent to $\mathbb{C}\mathrm{ov}_{[\cdot|\mathbf{Z}]}[\bar{\mathbf{\Psi}}_1,\bar{\mathbf{\Psi}}_2]$. And by symmetry, $(n-1)\sum_j(\mathrm{V})(\mathrm{II})^\mathsf{T}$ is equivalent to $\mathbb{C}\mathrm{ov}_{[\cdot|\mathbf{Z}]}[\bar{\mathbf{\Psi}}_2,\bar{\mathbf{\Psi}}_1]$. And as a short digression,

$$\begin{aligned}
(n-1)\sum_j (\mathrm{V})^2 &= \frac{1}{n-1}\sum_j\left(\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j-r_j\right)\right)^2 = \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + o_\mathbb{P}(1) \\
&= \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\left(\mathbf{b}_i-\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]+\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\right)\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + o_\mathbb{P}(1) \\
&= \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\left(\mathbf{b}_i-\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\right)\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \\
&\quad + \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\frac{\pi_{ij}}{1-\pi_{jj}}\right)\left(\sum_{i,i\neq j}\left(\mathbf{b}_i-\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\right)\frac{\pi_{ij}}{1-\pi_{jj}}\right)^\mathsf{T} + o_\mathbb{P}(1) \\
&= \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\pi_{ij}\right)^2 + \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\left(\mathbf{b}_i-\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\right)\pi_{ij}\right)^2 \\
&\quad + \frac{1}{n-1}\sum_j\varepsilon_j^2\left(\sum_{i,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\pi_{ij}\right)\left(\sum_{i,i\neq j}\left(\mathbf{b}_i-\mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\right)\pi_{ij}\right)^\mathsf{T} + o_\mathbb{P}(1),
\end{aligned}$$

where the first term in the above display recovers $\mathbb{V}_{[\cdot|\mathbf{Z}]}[\bar{\mathbf{\Psi}}_2]$, while the rest two are negligible by conditional expectation calculation. Therefore we recovered the asymptotic variance.

Back to the interaction terms,

$$\begin{aligned}
\left|(n-1)\sum_j (\mathrm{II})(\mathrm{VI})^\mathsf{T}\right| &= \left|\frac{1}{n-1}\sum_j \mathbf{a}_j\sum_{i,i\neq j}\mathbf{c}_i^\mathsf{T}\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j-r_j\right)^2\right| \\
&\precsim_\mathbb{P} \frac{1}{n-1}\sum_{i,j}\pi_{ij}^2 = o_\mathbb{P}(1),
\end{aligned}$$

and

$$\begin{aligned}
\Big|(n-1)\sum_j \text{(II)(VII)}^\mathsf{T}\Big| &= \Big|\frac{2}{n-1}\sum_j \mathbf{a}_j\Big(\hat{\mu}_j - r_j\Big)\sum_{i,i\neq j}\mathbf{c}_i^\mathsf{T}\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat{\mu}_i-\mu_i\Big)\Big| \\
&\asymp_p \Big|\frac{2}{n-1}\sum_{i,j}\mathbf{a}_j\Big(\hat{\mu}_j - r_j\Big)\mathbf{c}_i^\mathsf{T}\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat{\mu}_i-\mu_i\Big)\Big| \\
&\asymp_p \Big|\frac{2}{n-1}\sum_{i,j}\mathbf{a}_j\Big(\hat{\mu}_j - r_j\Big)\mathbf{c}_i^\mathsf{T}\pi_{ij}\Big(\hat{\mu}_i-\mu_i\Big)\Big| && \text{(Assumption A.3(2))} \\
&\leq \frac{2}{n-1}\cdot\sqrt{\sum_j |\mathbf{a}_j|^2\Big(\hat{\mu}_j - r_j\Big)^2}\sqrt{\sum_j |\mathbf{c}_j|^2\Big(\hat{\mu}_j-\mu_j\Big)^2} && \text{(Projection and Cauchy-Schwarz)} \\
&\leq o_\mathbb{P}(1)\cdot\frac{2}{n-1}\cdot\sqrt{\sum_j |\mathbf{a}_j|^2\Big(\hat{\mu}_j - r_j\Big)^2}\sqrt{\sum_j |\mathbf{c}_j|^2} = o_\mathbb{P}(1),
\end{aligned}$$

With a quick inspection, the above method also applies to the following interactions

$$(n-1)\sum_j \text{(III)(V)}^\mathsf{T} = o_\mathbb{P}(1), \qquad (n-1)\sum_j \text{(III)(VI)}^\mathsf{T} = o_\mathbb{P}(1), \qquad (n-1)\sum_j \text{(III)(VII)}^\mathsf{T} = o_\mathbb{P}(1),$$

and

$$(n-1)\sum_j \text{(IV)(V)}^\mathsf{T} = o_\mathbb{P}(1), \qquad (n-1)\sum_j \text{(IV)(VI)}^\mathsf{T} = o_\mathbb{P}(1), \qquad (n-1)\sum_j \text{(IV)(VII)}^\mathsf{T} = o_\mathbb{P}(1).$$

Next we consider the squared terms involving (III) and (IV):

$$(n-1)\sum_j \text{(III)}^2 = \frac{1}{n-1}\sum_j (\mathbf{b}_j)^2\Big(\hat{\mu}_j-\mu_j\Big)^2 \leq o_\mathbb{P}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{b}_j|^2 = o_\mathbb{P}(1),$$

and

$$(n-1)\sum_j \text{(IV)}^2 = \frac{1}{n-1}\sum_j (\mathbf{c}_j)^2\Big(\hat{\mu}_j-\mu_j\Big)^4 \leq o_\mathbb{P}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{c}_j|^2 = o_\mathbb{P}(1).$$

What remains are $\text{(V)(VI)}^\mathsf{T}$, $\text{(V)(VII)}^\mathsf{T}$, (VI)^2, $\text{(VI)(VII)}^\mathsf{T}$ and (VII)^2.

$$\begin{aligned}
\Big|(n-1)\sum_j \text{(V)(VI)}^\mathsf{T}\Big| &= \Big|\frac{1}{n-1}\sum_j\left(\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat{\mu}_j-r_j\Big)\right)\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\left(\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\right)^2\Big(\hat{\mu}_j-r_j\Big)^2\right)^\mathsf{T}\Big| \\
&\asymp_p \Big|\frac{1}{n-1}\sum_{i,j}\mathbf{b}_i\pi_{ij}\Big(\hat{\mu}_j-r_j\Big)^3\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\left(\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\right)^2\right)^\mathsf{T}\Big| \\
&\lesssim_\mathbb{P} \sqrt{\frac{1}{n-1}\sum_j\Big(\hat{\mu}_j-r_j\Big)^6\left|\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\left(\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\right)^2\right|^2} && \text{(Projection and Cauchy-Schwarz)} \\
&\asymp_p \sqrt{\frac{1}{n}\sum_{j,i,\ell}\pi_{ij}^2\pi_{\ell j}^2} = o_\mathbb{P}(1).
\end{aligned}$$

And

$$\begin{aligned}
\Big|(n-1)\sum_j \text{(V)(VII)}^\mathsf{T}\Big| &= \Big|\frac{2}{n-1}\sum_j\left(\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat{\mu}_j-r_j\Big)\right)\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\Big(\hat{\mu}_\ell-\mu_\ell\Big)\Big(\hat{\mu}_j-r_j\Big)\right)^\mathsf{T}\Big| \\
&\asymp_p \Big|\frac{1}{n-1}\sum_{i,j}\mathbf{b}_i\pi_{ij}\Big(\hat{\mu}_j-r_j\Big)^2\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\Big(\hat{\mu}_\ell-\mu_\ell\Big)\right)^\mathsf{T}\Big|
\end{aligned}$$

$$\lesssim_{\mathbb{P}} \sqrt{\frac{1}{n-1}\sum_j \left(\hat{\mu}_j - r_j\right)^4 \left|\sum_{\ell,\ell\neq j} \mathbf{c}_\ell \frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\left(\hat{\mu}_\ell - \mu_\ell\right)\right|^2} = o_{\mathbb{P}}(1),$$

(Projection and Cauchy-Schwarz)

where the last line uses Assumption A.1(3). Using techniques in the above results, we can show

$$(n-1)\sum_j (\mathrm{VI})^2 = o_{\mathbb{P}}(1), \qquad (n-1)\sum_j (\mathrm{VII})^2 = o_{\mathbb{P}}(1), \qquad (n-1)\sum_j (\mathrm{VI})(\mathrm{VII})^\mathsf{T} = o_{\mathbb{P}}(1),$$

which closes the proof. ■

SA-10.21 Lemma SA.18

Recall that we have the decomposition:

$$\begin{aligned}\max_{1\le i\le n} |\hat{\mu}_i^\star - \mu_i| &\le \max_{1\le i\le n} |\hat{\mu}_i^\star - \hat{\mu}_i| + \max_{1\le i\le n} |\hat{\mu}_i - \mu_i| \\ &\le \max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star \varepsilon_j| + \max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star (\hat{\mu}_j - \mu_j)| + \max_{1\le i\le n} |\hat{\mu}_i - \mu_i|.\end{aligned}$$

The last term is $o_{\mathbb{P}}(1)$ by Assumption A.1(3), and the second term can be bounded by using the conditional Hoeffding's inequality since $e_i^\star$ are bounded random variables with zero mean:

$$\begin{aligned}\mathbb{P}^\star\left[\max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star (\hat{\mu}_j - \mu_j)| \ge t\right] &\le n \cdot \max_{1\le i\le n} \mathbb{P}^\star\left[\max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star (\hat{\mu}_j - \mu_j)| \ge t\right] \\ &\le 2n \cdot \max_{1\le i\le n} \exp\left[-\frac{Ct^2}{\sum_j \pi_{ij}^2 |\hat{\mu}_j - \mu_j|^2}\right] \\ &\le 2n \cdot \exp\left[-\frac{Ct^2}{\left(\max_{1\le i\le n}\pi_{ii}\right)\left(\max_{1\le i\le n}|\hat{\mu}_i - \mu_i|^2\right)}\right],\end{aligned}$$

which goes to zero in probability if $\left(\max_{1\le i\le n}\pi_{ii}\right)\left(\max_{1\le i\le n}|\hat{\mu}_i - \mu_i|^2\right)/\log(n) = o_{\mathbb{P}}(1)$, or $\max_{1\le i\le n}\pi_{ii} = O_{\mathbb{P}}(1/\log(n))$. Since the conditional probability is always bounded by 1, the unconditional probability converges to zero by dominated convergence.

Next we consider the first term, which requires a reversed symmetrization inequality of (van der Vaart and Wellner, 1996, Lemma 2.3.7). And for simplicity, we assume $e_i^\star$ has Rademacher distribution, which is without loss of generality since we assumed $e_i^\star$ being bounded and symmetrically distributed with zero mean. Let ε_j' be an independent copy of ε_j from the conditional distribution $\varepsilon_j|\mathbf{z}_j, \mu_j$, then (subscript $[\cdot|\mathbf{Z}, \mathbf{e}^\star]$ indicates conditioning on $\mathbf{Z}$ and the bootstrap weights)

$$\alpha_n \cdot \mathbb{P}_{[\cdot|\mathbf{Z},\mathbf{e}^\star]}\left[\max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star \varepsilon_j| > t\right] \le \mathbb{P}_{[\cdot|\mathbf{Z},\mathbf{e}^\star]}\left[\max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star (\varepsilon_j - \varepsilon_j')| > \frac{t}{2}\right],$$

where

$$\begin{aligned}\alpha_n &= \min_{1\le i\le n} \mathbb{P}_{[\cdot|\mathbf{Z},\mathbf{e}^\star]}\left[|\sum_j \pi_{ij} e_j^\star \varepsilon_j| \le \frac{t}{2}\right] = 1 - \max_{1\le i\le n} \mathbb{P}_{[\cdot|\mathbf{Z},\mathbf{e}^\star]}\left[|\sum_j \pi_{ij} e_j^\star \varepsilon_j| > \frac{t}{2}\right] \\ &\ge 1 - \frac{Ct^2}{4}\max_{1\le i\le n}\pi_{ii},\end{aligned}$$

where $C \ge \max_{1\le i\le n}\mathbb{E}[\varepsilon_i^2|\mathbf{z}_i]$. Since we are dealing with probabilities, we can replace α_n by $\alpha_n^+ = \alpha_n \vee 0$ in the original inequality, yielding

$$\mathbb{P}_{[\cdot|\mathbf{Z},\mathbf{e}^\star]}\left[\max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star \varepsilon_j| > t\right] \le \frac{1}{\alpha_n^+}\mathbb{P}_{[\cdot|\mathbf{Z},\mathbf{e}^\star]}\left[\max_{1\le i\le n} |\sum_j \pi_{ij} e_j^\star (\varepsilon_j - \varepsilon_j')| > \frac{t}{2}\right].$$

Note that both sides in the above display are nonnegative, hence we can take expectation with respect to the bootstrap

weights,

$$\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}e_j^\star\varepsilon_j|>t\right] \le \frac{1}{\alpha_n^+}\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}(\varepsilon_j-\varepsilon_j')|>\frac{t}{2}\right]$$
$$\le \frac{2}{\alpha_n^+}\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}\varepsilon_j|>\frac{t}{4}\right],$$

where for the first line we used the fact that α_n^+ depends only on $\mathbf{Z}$ and both $e_j^\star$ and $\varepsilon_j-\varepsilon_j'$ have symmetric distribution. The second line is a simple fact of triangle inequality. Also note that the LHS in the above display is bounded by 1, hence we are able to tighten the RHS:

$$\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}e_j^\star\varepsilon_j|>t\right] \le \left(\frac{2}{\alpha_n^+}\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}\varepsilon_j|>\frac{t}{4}\right]\right)\wedge 1.$$

Now we go back to Assumptions A.1(5) and A.1(3). They jointly imply that

$$\max_{1\le i\le}|\sum_j \pi_{ij}\varepsilon_j| = o_{\mathbb{P}}(1) \qquad \Leftrightarrow \qquad \mathbb{E}\left[\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}\varepsilon_j>\frac{t}{4}\right]\right]\to 0,$$

which in turn implies that

$$\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}\varepsilon_j>\frac{t}{4}\right] = o_{\mathbb{P}}(1).$$

By our assumption, $\max_{1\le i\le n}\pi_{ii}=o_{\mathbb{P}}(1)$, hence $1/\alpha_n^+=O_{\mathbb{P}}(1)$, hence

$$\left(\frac{2}{\alpha_n^+}\mathbb{P}_{[\cdot|\mathbf{Z}]}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}\varepsilon_j|>\frac{t}{4}\right]\right)\wedge 1 = o_{\mathbb{P}}(1).$$

And since the quantity in the above display is bounded, dominated convergence implies

$$\mathbb{P}\left[\max_{1\le i\le n}|\sum_j \pi_{ij}e_j^\star\varepsilon_j|>t\right]\to 0,$$

which closes the proof. ■

SA-10.22 Proposition SA.19

The proof resembles that of Theorem SA.1, and is omitted here. ■

SA-10.23 Lemma SA.20

Note that

$$\begin{aligned}(\text{E.32}) &= \frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i,\hat\mu_i,\hat{\boldsymbol\theta})\\ &= \frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\mathbf{m}(\mathbf{w}_i,\hat\mu_i,\hat{\boldsymbol\theta})+o_{\mathbb{P}}(1)\\ &= \frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\mathbf{m}(\mathbf{w}_i,\hat\mu_i,\boldsymbol\theta_0)+o_{\mathbb{P}}(1).\end{aligned}$$

The second line uses (E.5), while the last line comes from the argument that

$$\frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\frac{\partial}{\partial\boldsymbol\theta}\mathbf{m}(\mathbf{w}_i,\hat\mu_i,\tilde{\boldsymbol\theta})\left(\hat{\boldsymbol\theta}-\boldsymbol\theta_0\right)\precsim_{\mathbb{P}}\frac{1}{n}\sum_i e_i^\star\cdot\frac{\partial}{\partial\boldsymbol\theta}\mathbf{m}(\mathbf{w}_i,\hat\mu_i,\tilde{\boldsymbol\theta})$$
$$\to_{\mathbb{P}}\mathbb{E}\left[e_i^\star\cdot\frac{\partial}{\partial\boldsymbol\theta}\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol\theta_0)\right],$$

given Assumption A.2(2). To further understand the last term, we still need to expand it with respect to $\hat{\mu}_i$, yielding

$$\frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \tag{I}$$

$$+ \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big) \tag{II}$$

$$+ \frac{1}{\sqrt{n}}\sum_i e_i^\star \cdot \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big)^2 \cdot (1 + o_{\mathbb{P}}(1)). \tag{III}$$

(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.4. 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.6 and SA.7. 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-10.24 Lemma SA.21

For (E.33), 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 $\boldsymbol{\theta}$:

$$\text{(E.33)} = \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big) + \frac{1}{n}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big),$$

where the second term is bounded by the following

$$\begin{aligned}
&\left|\frac{1}{n}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big)\right| \\
&\precsim_{\mathbb{P}} \frac{1}{n}\sum_i \left|\frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)\right| \\
&= o_{\mathbb{P}}(1) \cdot \frac{1}{n}\sum_i \left|\frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\right| = o_{\mathbb{P}}(1),
\end{aligned}$$

where the last line uses the uniform consistency of $\hat{\mu}_i^\star$ and $\hat{\mu}_i$. Hence

$$\begin{aligned}
\text{(E.33)} &= \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big) + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\sum_j \pi_{ij}\varepsilon_j e_j^\star\Big) \\
&- \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\sum_j \pi_{ij}(\hat{\mu}_j - \mu_j)e_j^\star\Big) + o_{\mathbb{P}}(1). \qquad \text{(I)}
\end{aligned}$$

For (I),

$$\begin{aligned}
\mathbb{E}^\star\left[(\text{I})(\text{I})^\mathsf{T}\right] &= \frac{1}{n}\mathbb{E}^\star\left[\sum_{i,i',j,j'} \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}^\star(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)^\mathsf{T}(\hat{\mu}_j - \mu_j)(\hat{\mu}_{j'} - \mu_{j'})e_j^\star e_{j'}^\star \pi_{ij}\pi_{i'j'}\right] \\
&= \frac{1}{n}\sum_{\substack{i,i',j \\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)^\mathsf{T}(\hat{\mu}_j - \mu_j)^2 \pi_{ij}\pi_{i'j} \qquad \text{(II)}
\end{aligned}$$

$$+\frac{2}{n}\sum_{\substack{i,i'\\ \text{distinct}}}\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_{i'},\hat{\mu}_{i'},\boldsymbol{\theta}_0)^\top(\hat{\mu}_i-\mu_i)(\hat{\mu}_{i'}-\mu_{i'})\pi_{ii}\pi_{i'i'} \tag{III}$$

$$+\frac{2}{n}\sum_{\substack{i,j\\ \text{distinct}}}\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)^\top(\hat{\mu}_j-\mu_j)^2\pi_{ij}^2 \tag{IV}$$

$$+\frac{C_1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_j,\hat{\mu}_j,\boldsymbol{\theta}_0)^\top(\hat{\mu}_j-\mu_j)^2\pi_{ij}\pi_{jj} \tag{V}$$

$$+\frac{C_2}{n}\sum_{\substack{i\\ \text{distinct}}}\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)^\top(\hat{\mu}_i-\mu_i)^2\pi_{ii}^2, \tag{VI}$$

where C_1 and C_2 are related to the third and fourth moments of $e_i^\star$. Then for each term,

$$\begin{aligned}
|(\mathrm{II})| &\le \left(\max_{1\le i\le n}|\hat{\mu}_i-\mu_i|^2\right)\cdot\frac{1}{n}\sum_{\substack{i,i'\\ \text{distinct}}}|\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)|\,|\dot{\mathbf{m}}(\mathbf{w}_{i'},\hat{\mu}_{i'},\boldsymbol{\theta}_0)|\,\pi_{ii'}\\
&\le o_{\mathbb{P}}(1)\cdot\frac{1}{n}\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)|^2 && \text{(projection and Assumption A.1(3))}\\
&= o_{\mathbb{P}}(1),
\end{aligned}$$

provided $\dot{\mathbf{m}}$ is Hölder continuous in μ_i. (III) can be handled by observing that

$$\begin{aligned}
|(\mathrm{III})| &\le \left(\frac{1}{\sqrt{n}}\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)|\,\pi_{ii}\,|\hat{\mu}_i-\mu_i|\right)^2\\
&\le o_{\mathbb{P}}(1)\cdot\left(\frac{1}{\sqrt{n}}\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\,\pi_{ii}\right)^2 = o_{\mathbb{P}}\left(\frac{k^2}{n}\right).
\end{aligned}$$

Similarly

$$|(\mathrm{IV})| \le o_{\mathbb{P}}(1)\cdot\frac{2}{n}\sum_{\substack{i,j\\ \text{distinct}}}|\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\,\pi_{ij}^2 = o_{\mathbb{P}}\left(\frac{k}{n}\right),$$

and

$$\begin{aligned}
|(\mathrm{V})| &\le \frac{C_1}{n}\left(\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)|^2\right)^{1/2}\left(\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)|^2\,|\hat{\mu}_i-\mu_i|^4\pi_{jj}^2\right)^{1/2}\\
&\precsim_{\mathbb{P}} n^{-1}\cdot\sqrt{n}\cdot\sqrt{k}\cdot o_{\mathbb{P}}(1) = o_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right).
\end{aligned}$$

Finally,

$$|(\mathrm{VI})| \le \frac{C_2}{n}\sum_i|\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)|^2\,|\hat{\mu}_i-\mu_i|^2\pi_{ii}^2 = o_{\mathbb{P}}\left(\frac{k}{n}\right).$$

To summarize, we have the following

$$\begin{aligned}
(\text{E.33}) &= \frac{1}{\sqrt{n}}\sum_i\dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0)\Big(\sum_j\pi_{ij}\varepsilon_j e_j^\star\Big)+o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}}\vee 1\right)\\
&= \frac{1}{\sqrt{n}}\sum_i\dot{\mathbf{m}}^\star(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\Big(\sum_j\pi_{ij}\varepsilon_j e_j^\star\Big)+o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}}\vee 1\right),
\end{aligned}$$

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.4, yielding

$$
\begin{aligned}
(\text{E.33}) = & \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0) | \mathbf{z}_j\right] \pi_{ij} \right) \varepsilon_i e_i^\star \\
& + \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}} \vee 1 \right).
\end{aligned}
$$

■

SA-10.25 Lemma SA.22

First note that

$$
\begin{aligned}
(\text{E.34}) &= \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&= \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \qquad \text{(I)} \\
&+ \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \left[\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) - \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \right] \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2,
\end{aligned}
$$

where the second term is easily bounded by

$$
\begin{aligned}
& \left| \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \left[\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) - \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \right] \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \right| \\
& \leq \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} (1 + e_i) \cdot \mathcal{H}^{\alpha,\delta}(\ddot{\mathbf{m}}_i) \cdot \left(|\tilde{\mu}_i^\star - \mu_i| + |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|\right)^\alpha \cdot |\hat{\mu}_i^\star - \hat{\mu}_i|^2 \\
& \leq o_{\mathbb{P}}(1) \cdot \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} (1 + e_i) \cdot \mathcal{H}^{\alpha,\delta}(\ddot{\mathbf{m}}_i) \cdot |\hat{\mu}_i^\star - \hat{\mu}_i|^2. \qquad \text{(II)}
\end{aligned}
$$

Compare (I) and (II) and note that Assumption A.2(4) imposes the same restrictions on $\ddot{\mathbf{m}}$ and $\mathcal{H}^{\alpha,\delta}(\ddot{\mathbf{m}}_i)$. 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 $\hat{\varepsilon}_\cdot$ by $\varepsilon_\cdot$. 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 $\mathbf{a}_i^\star = \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(\hat{\mu}_i^\star - \hat{\mu}_i)$)

$$
\begin{aligned}
\mathbb{E}^\star\left[(\text{III})(\text{III})^\mathsf{T}\right] &= \frac{1}{4n} \sum_{i,i',j,j'} \mathbb{E}^\star\left[\mathbf{a}_i^\star \mathbf{a}_{i'}^{\star\mathsf{T}} (\hat{\mu}_j - \mu_j)(\hat{\mu}_{j'} - \mu_{j'}) e_j^\star e_{j'}^\star \pi_{ij} \pi_{i'j'}\right] \\
&= \frac{1}{4n} \sum_{\substack{i,i',j \\ \text{distinct}}} \mathbb{E}^\star\left[\mathbf{a}_i^\star \mathbf{a}_{i'}^{\star\mathsf{T}}\right] (\hat{\mu}_j - \mu_j)^2 \pi_{ij} \pi_{i'j} \qquad \text{(IV)} \\
&+ \frac{1}{4n} \sum_{\substack{i,j \\ \text{distinct}}} \mathbb{E}^\star\left[\mathbf{a}_i^\star \mathbf{a}_i^{\star\mathsf{T}}\right] (\hat{\mu}_j - \mu_j)^2 \pi_{ij}^2 \qquad \text{(V)}
\end{aligned}
$$

$$+\frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star]\mathbb{E}^\star[\mathbf{a}_{i'}^\star e_{i'}^\star]^\mathsf{T}(\hat{\mu}_i-\mu_i)(\hat{\mu}_{i'}-\mu_{i'})\pi_{ii}\pi_{i'i'} \tag{VI}$$

$$+\frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]\mathbb{E}^\star\left[e_{i'}^{\star 2}\mathbf{a}_{i'}^{\star\mathsf{T}}\right](\hat{\mu}_{i'}-\mu_{i'})^2\pi_{ii'}\pi_{i'i'} \tag{VII}$$

$$+\frac{1}{4n}\sum_i\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_i^{\star\mathsf{T}}e_i^{\star 2}\right](\hat{\mu}_i-\mu_i)^2\pi_{ii}^2. \tag{VIII}$$

Then

$$\begin{aligned}
|(\mathrm{IV})| &= \left|\frac{1}{4n}\sum_{\substack{i,i',j\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathsf{T}}\right](\hat{\mu}_j-\mu_j)^2\pi_{ij}\pi_{i'j}\right|\\
&\precsim o_\mathbb{P}(1)\cdot\frac{1}{n}\sum_{i,i'}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathsf{T}}\right]\pi_{ii'}\\
&\le o_\mathbb{P}(1)\cdot\frac{1}{n}\sum_{i,i'}\mathbb{E}^\star\left[|\mathbf{a}_i^\star|\right]\mathbb{E}^\star\left[|\mathbf{a}_{i'}^\star|\right]\pi_{ii'}\\
&\le o_\mathbb{P}(1)\cdot\frac{1}{n}\sum_{i,i'}|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)||\ddot{\mathbf{m}}(\mathbf{w}_{i'},\mu_{i'},\boldsymbol{\theta}_0)|\pi_{ii'}\\
&\le o_\mathbb{P}(1)\cdot\frac{1}{n}\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2=o_\mathbb{P}(1),
\end{aligned}$$

where the second line uses Assumption A.1(3), the fourth line uses Assumption A.4(2), and the last line uses projection property and Assumption A.2(4). Similarly, we have, for (V),

$$\begin{aligned}
|(\mathrm{V})| &= \left|\frac{1}{4n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_i^{\star\mathsf{T}}\right](\hat{\mu}_j-\mu_j)^2\pi_{ij}^2\right|\\
&\precsim o_\mathbb{P}(1)\cdot\frac{1}{n}\sum_{i,j}|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\pi_{ij}^2=o_\mathbb{P}\left(\frac{k}{n}\right),
\end{aligned}$$

and the last equality is a simple consequence of Assumption A.2(4). (VI) is the most difficult, which can be rewritten as

$$\begin{aligned}
|(\mathrm{VI})| &= \frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star]\mathbb{E}^\star[\mathbf{a}_{i'}^\star e_{i'}^\star]^\mathsf{T}(\hat{\mu}_i-\mu_i)(\hat{\mu}_{i'}-\mu_{i'})\pi_{ii}\pi_{i'i'}\\
&\asymp\left(\frac{1}{\sqrt{n}}\sum_i\mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star](\hat{\mu}_i-\mu_i)\pi_{ii}\right)^2\\
&\precsim o_\mathbb{P}(1)\cdot\left(\frac{1}{\sqrt{n}}\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\pi_{ii}\right)^2=o_\mathbb{P}\left(\frac{k^2}{n}\right).
\end{aligned}$$

And

$$\begin{aligned}
|(\mathrm{VII})| &= \left|\frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]\mathbb{E}^\star\left[e_{i'}^{\star 2}\mathbf{a}_{i'}^{\star\mathsf{T}}\right](\hat{\mu}_{i'}-\mu_{i'})^2\pi_{ii'}\pi_{i'i'}\right|\\
&\precsim\frac{1}{n}\left(\sum_i|\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]|^2\right)^{1/2}\left(\sum_i|\mathbb{E}^\star\left[e_i^{\star 2}\mathbf{a}_i^\star\right]|^2|\hat{\mu}_i-\mu_i|^2\pi_{ii}^2\right)^{1/2} && \text{(projection)}\\
&\le o_\mathbb{P}(1)\cdot\frac{1}{n}\left(\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\right)^{1/2}\left(\sum_i|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2\pi_{ii}^2\right)^{1/2}
\end{aligned}$$

$$= o_{\mathbb{P}}(1) \cdot n^{-1} \cdot n^{1/2} \cdot k^{1/2} = o_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right).$$

Finally

$$\begin{aligned}|(\text{VII})| &= \frac{1}{4n}\sum_i \mathbb{E}^\star\left[\mathbf{a}_i^\star \mathbf{a}_i^{\star\mathsf{T}} e_i^{\star 2}\right](\hat{\mu}_i - \mu_i)^2 \pi_{ii}^2 \\ &\lesssim o_{\mathbb{P}}(1) \cdot \frac{1}{n}\sum_i |\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|^2 \pi_{ii}^2 = o_{\mathbb{P}}\left(\frac{k}{n}\right).\end{aligned}$$

Hence we have shown that

$$(\text{I}) = \frac{1}{\sqrt{n}}\sum_{i,j,\ell}\frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\hat{\varepsilon}_j \varepsilon_\ell e_j^\star e_\ell^\star \pi_{ij}\pi_{i\ell} + o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}} \vee 1\right).$$

Not surprisingly, we can replicate the above argument, and replace $\hat{\varepsilon}_j$ by ε_j in the above display, yielding

$$(\text{I}) = \frac{1}{\sqrt{n}}\sum_{i,j,\ell}\frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\varepsilon_j \varepsilon_\ell e_j^\star e_\ell^\star \pi_{ij}\pi_{i\ell} + o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}} \vee 1\right).$$

The next step is to apply Lemma SA.7 to conclude that

$$\begin{aligned}(\text{I}) &= \frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\varepsilon_j^2 e_j^{\star 2}\right]\pi_{ij}^2 + o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}} \vee 1\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{2,ii} \cdot \pi_{ii}^2 \cdot \mathbb{E}[e_i^{\star 3}] + o_{\mathbb{P}}\left(\frac{k}{\sqrt{n}} \vee 1\right).\end{aligned}$$

■

SA-10.26 Proposition SA.23

This is a simple consequence of linearization, Lemma SA.20, SA.21 and SA.22. ■

SA-10.27 Proposition SA.24, Part 1

For the ease of exposition we ignore (asymptotic negligible) remainder terms in the proof. Then $\hat{\boldsymbol{\theta}}^\star$ has the expansion

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}}\right) = \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^\star \hat{\mathbf{a}}_i + \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^\star \hat{\mathbf{b}}_i(\hat{\mu}_i^\star - \hat{\mu}_i) + \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^\star \hat{\mathbf{c}}_i(\hat{\mu}_i^\star - \hat{\mu}_i)^2,$$

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

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}\right) &= \frac{\sqrt{n}}{n_\omega(n_\omega - 1)} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{a}}_i \\ &\quad + \frac{\sqrt{n}}{n_\omega(n_\omega - 1)} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) + \frac{\sqrt{n}}{n_\omega(n_\omega - 1)} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\ &\quad + \frac{\sqrt{n}}{n_\omega(n_\omega - 1)} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1 - \pi_{jj}}(\hat{\mu}_j^\star - r_j^\star) \\ &\quad + \frac{2\sqrt{n}}{n_\omega(n_\omega - 1)} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \frac{\pi_{ij}}{1 - \pi_{jj}}(\hat{\mu}_i^\star - \hat{\mu}_i)(\hat{\mu}_j^\star - r_j^\star) \\ &\quad + \frac{\sqrt{n}}{n_\omega(n_\omega - 1)} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1 - \pi_{jj}}\right)^2 (\hat{\mu}_j^\star - r_j^\star)^2.\end{aligned}$$

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,

$$\begin{aligned}(n_\omega - 1)\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}^\star\right) &= \frac{\sqrt{n}}{n_\omega} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1 - \pi_{jj}}(\hat{\mu}_j^\star - r_j^\star) \\ &\quad + \frac{2\sqrt{n}}{n_\omega} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \frac{\pi_{ij}}{1 - \pi_{jj}}(\hat{\mu}_i^\star - \hat{\mu}_i)(\hat{\mu}_j^\star - r_j^\star)\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\sqrt{n}}{n_\omega} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 (\hat{\mu}_j^\star - r_j^\star)^2 \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} (\hat{\mu}_j^\star - r_j^\star) && \text{(I)} \\
&+ \frac{2}{\sqrt{n}} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}} (\hat{\mu}_i^\star - \hat{\mu}_i)(\hat{\mu}_j^\star - r_j^\star) && \text{(II)} \\
&+ \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 (\hat{\mu}_j^\star - r_j^\star)^2 && \text{(III)} \\
&+ o_{\mathbb{P}}(1).
\end{aligned}
$$

Next we analyze each term. For term (I), it is

$$
\begin{aligned}
\text{(I)} &= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} (\hat{\mu}_j^\star - r_j^\star) = \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \Big(\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell - e_j^\star \hat{\varepsilon}_j \Big) \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,\ell} \omega_j^\star(\omega_i^\star - \delta_{ij}) e_\ell^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{j\ell} \hat{\varepsilon}_\ell && \text{(I.1)} \\
&- \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star e_j^\star (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j. && \text{(I.2)}
\end{aligned}
$$

Again we consider conditional expectation:

$$
\begin{aligned}
\mathbb{E}^\star[(\text{I.1})] &= \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,\ell} \omega_j^\star(\omega_i^\star - \delta_{ij}) e_\ell^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{j\ell} \hat{\varepsilon}_\ell \right] \\
&= \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \right] + \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_j^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] \\
&+ \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_i \omega_i^\star(\omega_i^\star - 1) e_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ii}^2}{1-\pi_{ii}} \hat{\varepsilon}_i \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i + \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} \hat{\varepsilon}_j + \frac{1}{\sqrt{n}} \sum_i (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}^2}{1-\pi_{ii}} \hat{\varepsilon}_i.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}^\star[(\text{I.2})] &= \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star e_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] \\
&= \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star e_j^\star \omega_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] + \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_i \omega_i^\star e_i^\star (\omega_i^\star - 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \hat{\varepsilon}_i \right] \\
&= -\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j - \frac{1}{\sqrt{n}} \sum_i (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \hat{\varepsilon}_i.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\mathbb{E}^\star[(\text{I})] &= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i && \text{(I.3)} \\
&- \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \pi_{ij} \hat{\varepsilon}_j && \text{(I.4)} \\
&- \frac{1}{\sqrt{n}} \sum_i (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \pi_{ii} \hat{\varepsilon}_i. && \text{(I.5)}
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\text{(I.3)} &= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \varepsilon_i + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \left(\pi_{ij}^2 + \frac{\pi_{ij}^2 \pi_{jj}}{1-\pi_{jj}} \right) \varepsilon_i + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \pi_{ij}^2 \varepsilon_i + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_i \pi_{ij}^2 \varepsilon_i + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i} \mathbf{b}_i \pi_{ii} \varepsilon_i + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}} \sum_{i} \mathbb{E}[\mathbf{b}_i \varepsilon_i | \mathbf{z}_i] \pi_{ii} + o_{\mathbb{P}}(1) \\
&= \mathbf{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_{i} \mathbf{b}_{1,i} \pi_{ii} + o_{\mathbb{P}}(1).
\end{aligned}
$$

The second line follows from consistency and (E.36); the third line follows from Assumption A.3(2) and (E.37); the fourth line is a simple fact of Lemma SA.4. 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:

$$
\begin{aligned}
\text{(II)} &= \frac{2}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}} (\hat{\mu}_i^\star - \hat{\mu}_i)(\hat{\mu}_j^\star - r_j^\star) \\
&= \frac{2}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}} (\sum_{\ell} \pi_{i\ell} e_\ell^\star \hat{\varepsilon}_\ell)(\sum_{\ell} \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell - e_j^\star \hat{\varepsilon}_j) \\
&= \frac{2}{\sqrt{n}} \sum_{i,j,\ell,\ell'} \omega_j^\star (\omega_i^\star - \delta_{ij}) e_\ell^\star e_{\ell'}^\star \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{i\ell} \pi_{j\ell'} \hat{\varepsilon}_\ell \hat{\varepsilon}_{\ell'} && \text{(II.1)} \\
&- \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \omega_j^\star e_j^\star (\omega_i^\star - \delta_{ij}) e_\ell^\star \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{i\ell} \hat{\varepsilon}_\ell \hat{\varepsilon}_j. && \text{(II.2)}
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathbb{E}^\star[\text{(II.1)}] = \mathbb{E}^\star &\left[\frac{2}{\sqrt{n}} \sum_{i} \omega_i^\star (\omega_i^\star - 1) e_i^\star e_i^\star \hat{\mathbf{c}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \pi_{ii} \pi_{ii} \hat{\varepsilon}_i \hat{\varepsilon}_i \right] \\
+ \mathbb{E}^\star &\left[\frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_i^\star e_i^\star \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{ii} \pi_{ji} \hat{\varepsilon}_i \hat{\varepsilon}_i \right] \\
+ \mathbb{E}^\star &\left[\frac{2}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_j^\star e_j^\star \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{ij} \pi_{jj} \hat{\varepsilon}_j \hat{\varepsilon}_j \right] \\
+ \mathbb{E}^\star &\left[\frac{2}{\sqrt{n}} \sum_{i,\ell,i\neq \ell} \omega_i^\star (\omega_i^\star - 1) e_\ell^\star e_\ell^\star \hat{\mathbf{c}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \pi_{i\ell} \pi_{i\ell} \hat{\varepsilon}_\ell \hat{\varepsilon}_\ell \right]
\end{aligned}
$$

$$
\begin{aligned}
&+\mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_i^\star e_j^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ii}\pi_{jj}\hat{\varepsilon}_i\hat{\varepsilon}_j\right]\\
&+\mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_j^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ij}\pi_{ji}\hat{\varepsilon}_j\hat{\varepsilon}_i\right]\\
&+\mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\omega_j^\star(\omega_i^\star-\delta_{ij})e_\ell^\star e_\ell^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{i\ell}\pi_{j\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_\ell\right]\\
&=\frac{1}{\sqrt{n}}O_{\mathbb{P}}\left(\sum_i\pi_{ii}^3+\sum_{i,j}\pi_{ij}^2\pi_{ii}+\sum_{i,j}\pi_{ij}^2\pi_{jj}+\sum_{i,\ell}\pi_{i\ell}^2\pi_{ii}\right)\\
&+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2\\
&=\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j+\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2+o_{\mathbb{P}}(1),
\end{aligned}
$$

where the $o_{\mathbb{P}}(1)$ terms follows from (E.37) 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

$$
\begin{aligned}
\mathbb{E}^\star[(\text{II})]&=\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j && \text{(II.3)}\\
&+\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j && \text{(II.4)}\\
&+\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2 && \text{(II.5)}\\
&-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j && \text{(II.6)}\\
&-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2+o_{\mathbb{P}}(1). && \text{(II.7)}
\end{aligned}
$$

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.37) 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.37) 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) \\
&\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}\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'})}} \\
&\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}$$

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.7.

(III) has the following expansion:

$$
\begin{aligned}
(\mathrm{III}) &= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\hat{\mu}_j^\star - r_j^\star)^2 \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell - e_j^\star \hat{\varepsilon}_j)^2 \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell)^2 && \text{(III.1)} \\
&- \frac{2}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell) e_j^\star \hat{\varepsilon}_j && \text{(III.2)} \\
&+ \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 (e_j^\star \hat{\varepsilon}_j)^2. && \text{(III.3)}
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathbb{E}^\star[(\mathrm{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.37), (E.38) and Assumption A.3(1). Next

$$
\begin{aligned}
\mathbb{E}^\star[(\mathrm{III.2})] &= \mathbb{E}^\star \left[-\frac{2}{\sqrt{n}} \sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \left(\sum_\ell \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell \right) e_j^\star \hat{\varepsilon}_j \right] \\
&= \mathbb{E}^\star \left[-\frac{2}{\sqrt{n}} \sum_{i} \omega_i^\star e_i^\star (\omega_i^\star - 1) e_i^\star \hat{\mathbf{c}}_i \left(\frac{\pi_{ii}}{1-\pi_{ii}} \right)^2 \pi_{ii} \hat{\varepsilon}_i \hat{\varepsilon}_i \right]
\end{aligned}
$$

$$
\begin{aligned}
&+\mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star e_j^\star\omega_i^\star e_i^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ji}\hat{\varepsilon}_i\hat{\varepsilon}_j\right]\\
&+\mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star e_j^\star\omega_i^\star e_j^\star\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{jj}\hat{\varepsilon}_j\hat{\varepsilon}_j\right]\\
&=-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ij}\hat{\varepsilon}_i\hat{\varepsilon}_j+o_{\mathbb{P}}(1)=-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ij}\varepsilon_i\varepsilon_j+o_{\mathbb{P}}(1)\\
&=-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}[\mathbf{c}_i\varepsilon_i\varepsilon_j|\mathbf{z}_i,\mathbf{z}_j]\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ij}+o_{\mathbb{P}}(1)=o_{\mathbb{P}}(1).
\end{aligned}
$$

Finally

$$
\begin{aligned}
\mathbb{E}^\star[(\text{III.3})]&=\frac{1}{\sqrt{n}}\sum_{i,j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\hat{\varepsilon}_j^2+o_{\mathbb{P}}(1)=\frac{1}{\sqrt{n}}\sum_{i,j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\varepsilon_j^2+o_{\mathbb{P}}(1)\\
&=(\mathbb{E}^\star[e_i^{\star 3}]+1)\boldsymbol{\Sigma}_0\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{b}_{2,ij}\pi_{ij}^2+o_{\mathbb{P}}(1).
\end{aligned}
$$

Given the previous results,

$$
\begin{aligned}
&(n_\omega-1)\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)}-\hat{\boldsymbol{\theta}}^\star\right)=(1-\mathbb{E}^\star[e_i^{\star 3}])\boldsymbol{\Sigma}_0\frac{1}{\sqrt{n}}\left(\sum_i\mathbf{b}_{1,i}\pi_{ii}+\sum_{i,j}\mathbf{b}_{2,ij}\pi_{ij}^2\right)+o_{\mathbb{P}}(1)\\
&=(1-\mathbb{E}^\star[e_i^{\star 3}])\boldsymbol{\mathcal{B}}+o_{\mathbb{P}}(1).
\end{aligned}
$$

■

SA-10.28 Proposition SA.24, 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

$$
\begin{aligned}
\hat{\boldsymbol{\mathcal{V}}}^\star&=(n-1)\sum_j\left(\hat{\boldsymbol{\theta}}^{\star,(j)}-\hat{\boldsymbol{\theta}}^\star\right)^2-\frac{1}{n-1}\left(\hat{\boldsymbol{\mathcal{B}}}^\star\right)^2\\
&=(n-1)\sum_j\left(\hat{\boldsymbol{\theta}}^{\star,(j)}-\hat{\boldsymbol{\theta}}^\star\right)^2+O_{\mathbb{P}}\left(\frac{1}{n}\right).
\end{aligned}
$$

Next recall that

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(j)}-\hat{\boldsymbol{\theta}}&=\frac{1}{n_\omega-1}\sum_i(\omega_i^\star-\delta_{ij})\hat{\mathbf{a}}_i+\frac{1}{n_\omega-1}\sum_i(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)\\
&+\frac{1}{n_\omega-1}\sum_i(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)^2.
\end{aligned}
$$

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

$$\begin{aligned}
&\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right) = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i - \frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) - \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right),
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right)^2 = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i - \frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)^2 \\
&= \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&+ \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1}\sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&- \frac{1}{n_\omega - 1}\hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&+ \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right).
\end{aligned}$$

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

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}} &= \frac{1}{n_\omega}\sum_j \omega_j^\star \left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}\right) \\
&= \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{a}}_i \\
&- \frac{1}{n_\omega(n_\omega-1)}\sum_j \omega_j^\star \hat{\mathbf{a}}_j \\
&+ \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \\
&- \frac{1}{n_\omega(n_\omega-1)}\sum_j \omega_i^\star \hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) \\
&- \frac{1}{n_\omega(n_\omega-1)}\sum_{i,j} (\omega_i^\star - \delta_{ij})\omega_j^\star \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&+ \frac{1}{n_\omega - 1}\sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 \\
&- \frac{1}{n_\omega(n_\omega-1)}\sum_j \omega_j^\star \hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&+ \frac{1}{n_\omega(n_\omega-1)}\sum_{i,j} (\omega_i^\star - \delta_{ij})\omega_j^\star \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega(n_\omega-1)}\sum_{i,j} (\omega_i^\star - \delta_{ij})\omega_j^\star \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right),
\end{aligned}
$$

which means

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)} &= \frac{1}{n_\omega - 1}\left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n_\omega}\right) - \frac{1}{n_\omega - 1}\hat{\mathbf{a}}_j - \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) - \frac{1}{n_\omega - 1}\hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&- \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) + \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1}\left(\hat{\boldsymbol{\theta}}_{\mathtt{bc}}^\star - \hat{\boldsymbol{\theta}}\right) && \text{(I)} \\
&- \frac{1}{n_\omega - 1}\hat{\mathbf{a}}_j && \text{(II)} \\
&- \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) && \text{(III)} \\
&- \frac{1}{n_\omega - 1}\hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 && \text{(IV)} \\
&- \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) && \text{(V)} \\
&+ \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 && \text{(VI)} \\
&- \frac{2}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right). && \text{(VII)}
\end{aligned}
$$

Term (I) is the easiest:

$$
(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{I})^2 \asymp \left(\hat{\boldsymbol{\theta}}_{\mathtt{bc}}^\star - \hat{\boldsymbol{\theta}}\right)^2 = o_{\mathbb{P}}(1),
$$

by consistency. Similarly

$$(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{I})\Big((\mathrm{II}) + \cdots (\mathrm{VII})\Big)^\mathsf{T} = \Big(\hat{\boldsymbol{\theta}}_{\mathsf{bc}}^\star - \hat{\boldsymbol{\theta}}\Big)\sum_j \omega_j^\star\Big((\mathrm{II}) + \cdots (\mathrm{VII})\Big)^\mathsf{T} = o_\mathbb{P}(1).$$

Next

$$(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{II})^2 = \frac{1}{n_\omega - 1}\sum_j \omega_j^\star\,(\hat{\mathbf{a}}_j)^2 \to_\mathbb{P} \mathbb{V}[\bar{\boldsymbol{\Psi}}_1].$$

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

$$\begin{aligned}
(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{II})(\mathrm{V})^\mathsf{T} &= \frac{1}{n_\omega - 1}\sum_{i,j} \hat{\mathbf{a}}_j\hat{\mathbf{b}}_i^\mathsf{T}\omega_j^\star(\omega_i^\star - \delta_{ij})\left(\frac{\pi_{ij}}{1-\pi_{jj}}e_j^\star\hat{\varepsilon}_j\right)\\
&= \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j\omega_j^\star e_j^\star\hat{\varepsilon}_j\sum_i\left[\hat{\mathbf{b}}_i^\mathsf{T}(\omega_i^\star - \delta_{ij})\frac{\pi_{ij}}{1-\pi_{jj}}\right]\\
&= \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j\omega_j^\star e_j^\star\hat{\varepsilon}_j\sum_i\left[\hat{\mathbf{b}}_i^\mathsf{T}\pi_{ij}\right] && \text{(i)}\\
&+ \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j\omega_j^\star e_j^\star\hat{\varepsilon}_j\sum_i\left[\hat{\mathbf{b}}_i^\mathsf{T}\frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}}\right] && \text{(ii)}\\
&+ \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j\omega_j^\star e_j^\star\hat{\varepsilon}_j\sum_{i,i\neq j}\left[\hat{\mathbf{b}}_i^\mathsf{T}e_i^\star\frac{\pi_{ij}}{1-\pi_{jj}}\right] && \text{(iii)}\\
&+ \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j\omega_j^\star e_j^\star\hat{\varepsilon}_j\left[\hat{\mathbf{b}}_j^\mathsf{T}(e_j^\star - 1)\frac{\pi_{jj}}{1-\pi_{jj}}\right]. && \text{(iv)}
\end{aligned}$$

Then we have (i) $\to_\mathbb{P} \mathbb{C}\mathrm{ov}[\bar{\boldsymbol{\Psi}}_1, \bar{\boldsymbol{\Psi}}_2|\mathbf{Z}]$, and the other terms are asymptotically negligible. This essentially uses the same technique (conditional mean and variance calculation) used for Lemma SA.4 and SA.7, and we do not repeat here. By taking transpose, we have $(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{V})(\mathrm{II})^\mathsf{T} \to_\mathbb{P} \mathbb{C}\mathrm{ov}[\bar{\boldsymbol{\Psi}}_2, \bar{\boldsymbol{\Psi}}_1|\mathbf{Z}]$. Further,

$$\begin{aligned}
\Big|(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{II})(\mathrm{VI})^\mathsf{T}\Big| &= \Big|\frac{1}{n_\omega - 1}\sum_j \omega_j^\star\hat{\mathbf{a}}_j\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(e_j^\star\hat{\varepsilon}_j\right)^2\Big|\\
&\precsim_\mathbb{P} \frac{1}{n}\sum_{i,j}\pi_{ij}^2 = o_\mathbb{P}(1),
\end{aligned}$$

and

$$\begin{aligned}
\Big|(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{II})(\mathrm{VII})^\mathsf{T}\Big| &= \Big|\frac{2}{n_\omega - 1}\sum_j \omega_j^\star e_j^\star\hat{\varepsilon}_j\hat{\mathbf{a}}_j\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i\right)\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)\Big|\\
&\precsim_\mathbb{P} \frac{1}{n}\cdot\sqrt{\sum_j|\omega_j^\star e_j^\star\hat{\varepsilon}_j\hat{\mathbf{a}}_j|^2}\sqrt{\sum_j|(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i\right)|^2}\\
&= o_\mathbb{P}(1).
\end{aligned}$$

Due to uniform consistency of $\hat{\mu}_j^\star$, the following are easy to establish:

$$(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{III})^2 = o_\mathbb{P}(1) \qquad (n_\omega - 1)\sum_j \omega_j^\star(\mathrm{III})(\mathrm{IV})^\mathsf{T} = o_\mathbb{P}(1) \qquad (n_\omega - 1)\sum_j \omega_j^\star(\mathrm{III})(\mathrm{V})^\mathsf{T} = o_\mathbb{P}(1)$$

$$(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{III})(\mathrm{VI})^\mathsf{T} = o_\mathbb{P}(1) \qquad (n_\omega - 1)\sum_j \omega_j^\star(\mathrm{III})(\mathrm{VII})^\mathsf{T} = o_\mathbb{P}(1),$$

as well as

$$(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{IV})^2 = o_{\mathbb{P}}(1) \qquad (n_\omega - 1)\sum_j \omega_j^\star(\mathrm{IV})(\mathrm{V})^\mathsf{T} = o_{\mathbb{P}}(1) \qquad (n_\omega - 1)\sum_j \omega_j^\star(\mathrm{IV})(\mathrm{VI})^\mathsf{T} = o_{\mathbb{P}}(1)$$
$$(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{IV})(\mathrm{VII})^\mathsf{T} = o_{\mathbb{P}}(1).$$

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 $(\mathrm{VII})^2$.

$$\begin{aligned}
&\left|(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{V})(\mathrm{VI})^\mathsf{T}\right| \\
&= \left|\frac{1}{n_\omega - 1}\sum_j \omega_j^\star \left(\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)\right)\left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j})\hat{\mathbf{c}}_\ell \left(\frac{\pi_{\ell j}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2\right)^\mathsf{T}\right| \\
&\asymp_{\mathbb{P}} \left|\frac{1}{n_\omega - 1}\sum_{i,j} \omega_j^\star (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(e_j^\star \hat{\varepsilon}_j\right)^3 \pi_{ij} \left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j})\hat{\mathbf{c}}_\ell \left(\frac{\pi_{\ell j}}{1-\pi_{jj}}\right)^2\right)^\mathsf{T}\right| \\
&\precsim_{\mathbb{P}} \left(\frac{1}{n}\sum_j \left|\sum_\ell (\omega_\ell^\star - \delta_{\ell j})\hat{\mathbf{c}}_\ell \left(\frac{\pi_{\ell j}}{1-\pi_{jj}}\right)^2\right|^2\right)^{1/2} \\
&\asymp_p \sqrt{\frac{1}{n}\sum_{j,i,\ell} \pi_{ij}^2 \pi_{\ell j}^2} = o_{\mathbb{P}}(1).
\end{aligned}$$

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. ■

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231–263.

——— (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72(1), 1–19.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186(2), 345–366.

Björklund, A., and R. Moffitt (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69(1), 42–49.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101(6), 2754–2781.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2017): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, forthcoming.

——— (2018): "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity," *Journal of the American Statistical Association*, forthcoming.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

Heckman, J. J., and E. J. Vytlacil (2005): "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Imbens, G. W., J. D. Angrist, and A. B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14(1).

Mammen, E. (1989): "Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap," *Annals of Statistics*, 17(1), 382–400.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–82.

Olley, G. S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64(6), 1263–1297.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer, New York.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2 edn.

Wooldridge, J. 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) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\mathtt{MTE,bc}} - \tau_{\mathtt{MTE}})$						
	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) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\mathtt{MTE,bc}} - \tau_{\mathtt{MTE}})$						
	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) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\mathtt{MTE,bc}} - \tau_{\mathtt{MTE}})$						
	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) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$							$\sqrt{n}(\hat{\tau}_{\mathtt{MTE,bc}} - \tau_{\mathtt{MTE}})$						
	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) $n = 1000$

			$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$: **conventional**							$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$: **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) $n = 2000$

			$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$: **conventional**							$\sqrt{n}(\hat{\tau}_{\mathtt{MTE}} - \tau_{\mathtt{MTE}})$: **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 ($p = 2$)

	$\hat{\tau}_{\mathtt{MTE}}(0.2)$					$\hat{\tau}_{\mathtt{MTE}}(0.5)$					$\hat{\tau}_{\mathtt{MTE}}(0.8)$				
	(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).