Two-Step Estimation and Inference
with Possibly Many Included Covariates[1]

MATIAS D. CATTANEO

Department of Economics and Department of Statistics, University of Michigan

MICHAEL JANSSON

Department of Economics, UC Berkeley and CREATES

XINWEI MA

Department of Economics, University of Michigan

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.

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 result is not specific to our simulation setting, as our general results apply broadly to many other treatment effect, policy evaluation, and applied microeconomics settings: 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 covariates/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 covariates 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 the jackknife bias correction and variance estimation deliver 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 et al. (2018) 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

---

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. The Supplemental Appendix discusses several examples employing two-step estimators with possibly many covariates.

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 on 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 building on the work of Carneiro et al. (2011). Finally, Section 8 concludes. The 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 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 parameter choices and underlying assumptions (e.g.,

smoothness). In particular, first, Cattaneo et al. (2013) and Cattaneo and Jansson (2018) 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 et al. (2018) 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 literature by offering new inference procedures 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 et al., 2013; Cattaneo et al., 2018b; Li and Müller, 2017, and references therein) by developing generic distributional approximations for two-step estimators where the first-step estimator is possibly high-dimensional. See also Fan et al. (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

---

2. A certain class of *linear* semiparametric estimators has a very different behavior when undersmoothing the first step nonparametric estimator; see Cattaneo et al. (2010, 2014b,a) and Cattaneo et al. (2018a) for discussion and references. In particular, their results 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).

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 of available covariates is larger than the sample size.

(e.g., Belloni et al., 2014; Farrell, 2015; Belloni et al., 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 a 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 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 (Fernandez-Val and Weidner, 2018, and references therein) 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 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 et al., 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., 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. 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 by projecting r_i on the possibly high-dimensional covariate $\mathbf{z}_i$ with least squares, as discussed further below, we study the two-step estimator:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \left| \boldsymbol{\Omega}_n^{1/2} \sum_{i=1}^n \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}) \right|, \tag{2.1}$$

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

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:

$$\mathcal{V}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \mathscr{B}) \rightsquigarrow \operatorname{Normal}(\mathbf{0}, \mathbf{I}), \tag{2.2}$$

where $\rightsquigarrow$ denotes convergence in distribution, with limits always taken as $n \to \infty$ and $k = O(\sqrt{n})$, and $\mathcal{V}$ 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 $\mathcal{V}^{-1/2}\mathscr{B} = O_{\mathbb{P}}(k/\sqrt{n})$ 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 (2.2) is the characterization and estimation of the variance $\mathcal{V}$. 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 $\mathcal{V}$ is harder to characterize and estimate. In fact, our distributional approximation leading to (2.2) 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 (2.2) 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:

$$\mathscr{T} \overset{\text{def}}{=} \hat{\mathscr{V}}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \hat{\mathscr{B}}) \rightsquigarrow \text{Normal}(\mathbf{0}, \mathbf{I}), \tag{2.3}$$

To implement the jackknife method, one first constructs $\hat{\boldsymbol{\theta}}^{(\ell)}$, for which the ℓth observation is deleted and then both steps are re-estimated using the remaining observations. Denote by $\hat{\boldsymbol{\theta}}^{(\cdot)} = n^{-1}\sum_{\ell=1}^{n}\hat{\boldsymbol{\theta}}^{(\ell)}$ the average of the leave-one-observation-out estimators, then

$$\hat{\mathscr{B}} = (n-1)(\hat{\boldsymbol{\theta}}^{(\cdot)} - \hat{\boldsymbol{\theta}}), \quad \hat{\mathscr{V}} = \frac{n-1}{n}\sum_{\ell=1}^{n}(\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\cdot)})(\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\cdot)})^{\mathsf{T}}. \tag{2.4}$$

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, a method that is particularly amenable to jackknifing.

While result (2.3) 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:

$$\sup_{t \in \mathbb{R}^{d_\theta}} \left| \mathbb{P}[\mathscr{T} \leq t] - \mathbb{P}^{\star}[\mathscr{T}^{\star} \leq t] \right| \to_{\mathbb{P}} 0, \qquad \mathscr{T}^{\star} \overset{\text{def}}{=} \hat{\mathscr{V}}^{\star^{-1/2}}(\hat{\boldsymbol{\theta}}^{\star} - \hat{\boldsymbol{\theta}} - \hat{\mathscr{B}}^{\star}), \tag{2.5}$$

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 possibly many covariates entering the first-step estimate can be conducted by combining results (2.3) and (2.5). 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 (2.5) 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 (2.1), and show that under fairly general 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 (2.2).

3.1. Regularity Conditions

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. In addition, define the transformation

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

The following assumption collects some basic notation and regularity conditions.

Assumption 1 (Regularity Conditions). *Let $0 < \delta,\ \alpha,\ C < \infty$ be some fixed constants.*

(i) $\mathbf{m}$ is twice continuously differentiable in μ with derivatives denoted by $\dot{\mathbf{m}}(\mathbf{w}_i,\mu,\boldsymbol{\theta}_0) = \frac{\partial}{\partial\mu}\mathbf{m}(\mathbf{w}_i,\mu,\boldsymbol{\theta}_0)$ and $\ddot{\mathbf{m}}(\mathbf{w}_i,\mu,\boldsymbol{\theta}_0) = \frac{\partial^2}{\partial\mu^2}\mathbf{m}(\mathbf{w}_i,\mu,\boldsymbol{\theta}_0)$. In addition, $\mathbf{m}$ and $\dot{\mathbf{m}}$ are continuously differentiable in $\boldsymbol{\theta}$.

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

(iii) $\mathbf{m}_i,\ \dot{\mathbf{m}}_i,\ \ddot{\mathbf{m}}_i,\ \mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}}),\ \varepsilon_i^3,\ |\dot{\mathbf{m}}_i\varepsilon_i|,\ |\ddot{\mathbf{m}}_i|\varepsilon_i^2,\ |\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}})|\varepsilon_i^2 \in \mathsf{BCM}_2$, where $\mathbf{m}_i = \mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta})$, $\dot{\mathbf{m}}_i = \dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta})$, $\ddot{\mathbf{m}}_i = \ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta})$, and $\varepsilon_i = r_i - \mu_i$.

(iv) $\mathbf{M}_0 = \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\right]$ has full (column) rank d_θ.

These conditions are standard in the literature. They require smoothness of $\mathbf{m}(\mathbf{w},\mu,\boldsymbol{\theta})$ with respect to both μ and $\boldsymbol{\theta}$, and boundedness of (conditional) moments of various orders. In future work we plan to extend our results to non-differentiable second-step estimating equations.

3.2. First-Step Estimation

We are interested in understanding the effects of including 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:

$$\mu_i = \mathbb{E}[r_i|\mathbf{z}_i] = \mathbf{z}_i^\mathsf{T}\boldsymbol{\beta} + \eta_i, \qquad \mathbb{E}[\mathbf{z}_i\eta_i] = \mathbf{0}, \tag{3.6}$$

for a non-random vector $\boldsymbol{\beta}$, where η_i represents the error in the best linear approximation. This motivates the least-squares first-step estimate:

$$\hat{\mu}_i = \mathbf{z}_i^\mathsf{T}\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^n (r_i - \mathbf{z}_i^\mathsf{T}\boldsymbol{\beta})^2, \tag{3.7}$$

which 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 (3.7), we investigate the implications of introducing possibly many covariates $\mathbf{z}_i$, and thus our approximations allow for (but do not require that) k being "large" relative to the sample size. Specifically, we show that when $k \propto \sqrt{n}$ conventional inference procedures become invalid due to a new bias term in the asymptotic approximations.

In some settings, the covariates $\mathbf{z}_i$ can have approximation power beyond the first-step estimation, as it occurs for instance when these covariates are basis expansions. To allow for this possibility, we also define, for a non-random matrix $\mathbf{\Gamma}$,

$$\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i] = \mathbf{\Gamma}\mathbf{z}_i + \boldsymbol{\zeta}_i, \qquad \mathbb{E}[\mathbf{z}_i\boldsymbol{\zeta}_i^{\mathsf{T}}] = \mathbf{0}, \qquad (3.8)$$

where $\boldsymbol{\zeta}_i$ is the error from the best linear approximation of $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ based on $\mathbf{z}_i$. This approximation error will not be small in general, because our paper allows for generic high-dimensional first-step covariates. However, in some special cases it can be small as we discuss further below.

The following assumption collects the key restrictions we impose on the first-step procedure.

Assumption 2 (First-Step).
(i) $\max_{1 \leq i \leq n} |\hat{\mu}_i - \mu_i| = o_{\mathbb{P}}(1)$.
(ii) $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$ *and* $\mathbb{E}[|\eta_i|^2]\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(n^{-1})$.

This assumption imposes high-level conditions on the covariates $\mathbf{z}_i$ entering the first-step estimate (3.7), covering both series-based nonparametric estimation and, more generally, many covariates settings. Assumption 2(i) requires uniform consistency of $\hat{\mu}_i$ for μ_i only, without a convergence rate. In Section SA-2 of the Supplemental Appendix we discuss primitive conditions in different scenarios, covering (i) nonparametric series-based methods (Belloni et al., 2015; Cattaneo et al., 2018), (ii) generic covariates with alternative conditions on the tails of their distribution, and (iii) generic covariates formed using many dummy/discrete regressors. The assumption also holds easily when covariates are discrete and a fully saturated model is used. This 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. Underlying this assumption is the implicit requirement that the best linear approximation of μ_i based on $\mathbf{z}_i$ in (3.6) should vanish asymptotically.

Assumption 2(ii) concerns the approximation power of the covariates $\mathbf{z}_i$ explicitly, measured in terms of the mean squared error of best linear approximations. It requires, at least, that the best linear approximation error in (3.6) is sufficiently small relative to the sample size in mean square. The condition $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$ cannot be dropped without affecting the interpretation of the final estimand $\boldsymbol{\theta}_0$ because the first-step best linear approximation error 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 latter 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.

Furthermore, the exact quality of approximation for the first-step estimate required in Assumption 2(ii) depends on the quality of approximation in (3.8). At one extreme, the covariates $\mathbf{z}_i$ may not offer any approximation of $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ in mean square, in which case $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = O(1)$, and hence the relevant restriction becomes $\mathbb{E}[|\eta_i|^2] = o(n^{-1})$. This corresponds to the case of many generic covariates $\mathbf{z}_i$ and non-linear $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$, that is, cases where $\mathbf{z}_i$ are not basis of approximation and/or $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ can not be well approximated by a linear combination of $\mathbf{z}_i$.

At the other extreme, if $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ can be well approximated by the best linear mean square prediction based on $\mathbf{z}_i$ so that, at least, $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = O(n^{-1/2})$, then the relevant restriction on the first-step estimate becomes $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$. This case encompasses the standard two-step semiparametric setup, where the covariates $\mathbf{z}_i$ include basis expansions able to approximate $\mu_i = \mathbb{E}[r_i|\mathbf{z}_i]$ and $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ accurately enough in mean square (usually justified by smoothness of these conditional expectations). From this perspective, the sufficient conditions $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$ and $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = O(n^{-1/2})$ reassemble the usual requirement of better than $n^{1/4}$-consistency of first-step nonparametric estimators in two-step semiparametrics (see Cattaneo and Jansson, 2018, and references therein), but this is imposed only on best linear approximation errors (i.e., misspecification/smoothing bias), which are exacerbated for small k and not for large k, the latter being the main focus of the present paper.

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 (3.6). 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 a 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.

3.3. Distribution Theory

It is not difficult to establish $\hat{\boldsymbol{\theta}} \to_{\mathbb{P}} \boldsymbol{\theta}_0$, even when $k/\sqrt{n} = O(1)$. Thus, we impose the following high-level assumption (see the Supplemental Appendix for lower-level conditions).

Assumption 3 (Consistency).
(i) $\hat{\boldsymbol{\theta}} \to_{\mathbb{P}} \boldsymbol{\theta}_0$ the unique solution of $\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})] = \mathbf{0}$ and an interior point of Θ.
(ii) $\Omega_n \to_{\mathbb{P}} \Omega_0$ positive definite.

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 "large" relative to the sample size. To describe heuristically the result, consistency of $\hat{\boldsymbol{\theta}}$ and a second-order

Taylor series expansion give:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx \quad \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \tag{3.9}$$

$$+ \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \left(\hat{\mu}_i - \mu_i \right) \tag{3.10}$$

$$+ \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \frac{1}{2} \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \left(\hat{\mu}_i - \mu_i \right)^2, \tag{3.11}$$

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

Term (3.9) 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 (3.10) contributes to the variability of $\hat{\boldsymbol{\theta}}$ as a result of estimating the first step, and term (3.11) will be negligible. Here, however, we show that under the many covariates assumption $k/\sqrt{n} \not\to 0$, both (3.10) and (3.11) 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 $\hat{\mu}_i - \mu_i$ also increases and features in terms (3.10) and (3.11). This, in turn, affects the finite sample performance of the usual asymptotic approximations, delivering unreliable results in applications. To be specific, the term (3.10) 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 (3.11) contributes with a bias arising from averaging (non-linear) squared errors in the estimation 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 $\boldsymbol{\Pi} = \mathbf{Z}(\mathbf{Z}^\mathsf{T} \mathbf{Z})^- \mathbf{Z}^\mathsf{T}$ be the projection matrix with elements $\{\pi_{ij} : 1 \leq i, j \leq n\}$.

Theorem 1 (Asymptotic Normality).
Suppose Assumptions 1, 2 and 3 hold. If $k = O(\sqrt{n})$, then (2.2) holds with

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

where

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

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

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 $\mathscr{V}^{-1/2}\mathscr{B}$ 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 (3.10), 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 $\mathbb{C}\text{ov}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0), r_i - \mu_i | \mathbf{z}_i]\pi_{ii}$. 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., $\mathbb{C}\text{ov}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0), r_i - \mu_i | \mathbf{z}_i] = \mathbf{0}$).

The second term in $\mathbf{B}_i$ captures the quadratic dependence of the estimating equation on the unobserved μ_i, coming from (3.11). 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 to the first step on average to the second order.

The first bias can 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 et al. (2018) introduced the class of locally robust estimators, which are a generalization of doubly robust estimators (e.g., Bang and Robins, 2005) and the efficient influence function estimators (e.g., 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 et al. (2018) and Newey and Robins (2018) for related approaches. This type of estimators are carefully constructed so that (3.10) is removed, but they do not account for (3.11). Because the "large" k bias is in part characterized by (3.11), 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. Cattaneo and Jansson (2018) illustrate this fact in the context of kernel-based estimation.

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

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

hence $\hat{\boldsymbol{\theta}}$ is asymptotically linear after bias correction even when $k/\sqrt{n} \nrightarrow 0$. 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 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 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,

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

provided that $\dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)$ is full rank. Hence, the usual delta method can be used for estimation and inference in our setting, despite the presence of potentially many covariates entering the 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.

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 (2.4). This approach is fully data-driven and automatic. In addition, another appealing feature of the jackknife in our case is that it is possible to exploit the specific structure of the problem to reduce computational 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 $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\mathsf{T}\mathbf{Z})^-\mathbf{Z}^\mathsf{T}$. 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 the following additional mild assumptions on the possibly large dimensional covariates $\mathbf{z}_i$, captured through the projection matrix of the first-step estimate.

Assumption 4 (Jackknife).
(i) $\sum_{1 \leq i \leq n} \pi_{ii}^2 = o_\mathbb{P}(k)$.
(ii) $\max_{1 \leq i \leq n} 1/(1 - \pi_{ii}) = O_\mathbb{P}(1)$.

The first 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). With these conditions, we obtain the following result.

Theorem 2 (Jackknife-Based Valid Inference).
Suppose Assumptions 1, 2, 3 and 4 hold. If $k = O(\sqrt{n})$, then (2.3) holds.

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.

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.

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 (2.3). 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 $\omega_i^\star$, $i = 1, 2, \cdots, n$ be i.i.d. bootstrap weights with $\mathbb{E}[\omega_i^\star] = 1$, $\mathbb{V}[\omega_i^\star] = 1$, $\mathbb{E}[(\omega_i^\star - 1)^3] = 0$ and finite fourth moment. 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 (3.7): 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 (2.1):

$$\hat{\boldsymbol{\theta}}^\star = \arg\min_{\boldsymbol{\theta}} \left| \boldsymbol{\Omega}_n^{1/2} \sum_{i=1}^n \omega_i^\star \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^\star, \boldsymbol{\theta}) \right|. \tag{5.12}$$

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

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

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 utilize 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.

Only two additional mild, high-level conditions on the bootstrap analogue first-step and second-step estimators are imposed as follows.

Assumption 5 (Bootstrap).
(i) $\max_{1 \leq i \leq n} |\hat{\mu}_i^\star - \hat{\mu}_i| = o_{\mathbb{P}}(1)$.
(ii) $|\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}}| = o_{\mathbb{P}}(1)$.

The following theorem summarizes our main result for inference employing the bootstrap after jackknife bias and variance estimation.

Theorem 3 (Bootstrap Validity).
Suppose Assumptions 1, 2, 3, 4 and 5 hold. If $k = O(\sqrt{n})$, then (2.5) 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 $\hat{q}_\alpha = \inf\{t \in \mathbb{R} : \hat{F}(t) \geq \alpha\}$ is the empirical α^{th} quantile of $\{\mathscr{T}_b^\star : 1 \leq b \leq B\}$, with $\hat{F}(t) = \frac{1}{B} \sum_{b=1}^B \mathbb{1}[\mathscr{T}_b^\star \leq t]$ and $\mathscr{T}_b^\star$ denoting the bootstrap statistic in (2.5) in b^{th} simulation.

6. EXAMPLES

We now apply our results to two examples: production function estimation, which is an IO application with multidimensional partially linear first step estimation, and marginal treatment effect estimation, which relates to IV methods with heterogeneous treatment effects and many covariates/instruments. The Supplemental Appendix analyzes several other examples 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 et al. (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:

$$Y_{i,t} = \beta_L L_{i,t} + \beta_K K_{i,t} + \beta_A A_{i,t} + W_{i,t} + U_{i,t}, \tag{6.13}$$

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 the right-hand-side variables). Since the productivity factor W is unobserved, (6.13) 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 writing $W_{i,t} = h_t(I_{i,t}, K_{i,t}, A_{i,t})$, with h_t unknown and possibly time-dependent. Then, (6.13) can be written as

$$Y_{i,t} = \beta_L L_{i,t} + \phi_{i,t} + U_{i,t}, \quad \phi_{i,t} = \beta_K K_{i,t} + \beta_A A_{i,t} + h_t(I_{i,t}, K_{i,t}, A_{i,t}). \quad (6.14)$$

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, Olley and Pakes (1996) use information embedded in the firm's exit decision, which is shown 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

$$P_{i,t} = P_t(I_{i,t}, K_{i,t}, A_{i,t}) = \mathbb{E}[\chi_{i,t+1}|I_{i,t}, K_{i,t}, A_{i,t}], \quad (6.15)$$

is a valid proxy for $\underline{W}_{i,t}$. Therefore, with the time index progressed by one period, we rewrite (6.13) as

$$Y_{i,t+1} - \beta_L L_{i,t+1} = \beta_K K_{i,t+1} + \beta_A A_{i,t+1} + g(P_{i,t}, \ W_{i,t}) + V_{i,t+1} + U_{i,t+1}$$
$$= \beta_K K_{i,t+1} + \beta_A A_{i,t+1} + g(P_{i,t}, \ \phi_{i,t} - \beta_K K_{i,t} - \beta_A A_{i,t}) + V_{i,t+1} + U_{i,t+1}. \ (6.16)$$

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 (6.16), 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 (6.14) 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 (6.15) 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:

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

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

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

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,

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

where

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

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

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

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

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

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

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

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 $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Xi}_0$ 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 $\boldsymbol{\Psi}_{i,2}$ 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 et al. (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 et al., 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,

$$Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0), \quad Y_i(1) = g_1(\mathbf{X}_i) + U_{1i}, \ Y_i(0) = g_0(\mathbf{X}_i) + U_{0i}, \quad (6.17)$$

$$T_i = \mathbb{1}[P_i \geq V_i], \quad P_i = P(\mathbf{Z}_i) = \mathbb{E}[T_i|\mathbf{Z}_i], \ V_i|\mathbf{X}_i \sim \text{Uniform}[0,1], \quad (6.18)$$

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 (6.18) is taken essentially without loss of generality to be of the single threshold-crossing form (see Vytlacil, 2002, for more discussion), though this representation may affect the interpretation of the unobserved heterogeneity.

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

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

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 (6.17) are unrelated to that of the selection equation (6.18). The parameter $\tau_{\text{MTE}}(a|\mathbf{x})$ is understood as 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, $0 < \mathbb{P}[T_i = 1|\mathbf{X}_i] < 1$, so that conditional on the covariates, both treated and untreated individuals are observable in the population. It can then be shown that, for any limit point a in the support of the propensity score, $\tau_{\text{MTE}}(a|\mathbf{x})$ is

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

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 $\tau_{\text{MTE}}(a|\mathbf{x})$ 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:

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

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

Identification and estimation of the MTE, as well as other policy-relevant parameters based on it, require exogenous variation in the treatment equation (6.18) 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 $\hat{\boldsymbol{\theta}}$,

$$\mathbf{B}_i = \frac{\partial^2 e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial P_i \partial \boldsymbol{\theta}} \Big[(1 - P_i) \cdot \mathbb{E}[T_i Y_i(1)|\mathbf{Z}_i] - P_i \cdot \mathbb{E}[(1 - T_i)Y_i(0)|\mathbf{Z}_i] \Big] \pi_{ii}$$

$$+ \frac{1}{2} \sum_{j=1}^{n} \left[\frac{\partial^2 e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial P_i \partial \boldsymbol{\theta}} \tau_{\mathtt{MTE}}(P_i|\mathbf{X}_i) + \frac{1}{2} \frac{\partial e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \tau_{\mathtt{MTE}}(P_i|\mathbf{X}_i)}{\partial P_i} \right] P_j(1 - P_j)\pi_{ij}^2,$$

$$\boldsymbol{\Psi}_i = \frac{\partial e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0) \Big) - \left(\sum_{j=1}^{n} \frac{\partial e(\mathbf{X}_j, P_j, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \tau_{\mathtt{MTE}}(P_j|\mathbf{X}_j)\pi_{ij} \right)(T_i - P_i),$$

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, $\hat{\tau}_{\mathtt{MTE}}(a|\mathbf{x})$, 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 that 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 negligible.

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 2 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 $\tau_{\text{MTE}}(a) = 1 - a$. 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 $\hat{\tau}_{\text{MTE}}(a) = \hat{\theta}_2 + 2a \cdot \hat{\theta}_3$ 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 $\sqrt{n}\,(\hat{\tau}_{\text{MTE}}(a) - \tau_{\text{MTE}}(a))$ 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 $(\hat{\tau}_{\text{MTE}} - \tau_{\text{MTE}})/\sqrt{\mathbb{V}[\hat{\tau}_{\text{MTE}}]}$, with $\mathbb{V}[\hat{\tau}_{\text{MTE}}]$ denoting the simulation variance of $\hat{\tau}_{\text{MTE}}$, 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 $(\hat{\tau}_{\text{MTE}} - \hat{\mathscr{B}} - \tau_{\text{MTE}})/\sqrt{\hat{\mathscr{V}}}$ 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 and **??**. 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 et al. (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 the 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, $\tau_{\text{MTE}}(a|\mathbf{x}) = \partial \mathbb{E}[Y_i | P_i = a, \mathbf{X}_i = \mathbf{x}]/\partial a$ 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 the 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 2 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 $\phi(a)$ 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 $\phi(\hat{P}_i) = [\hat{P}_i, \hat{P}_i^2, \hat{P}_i^3, \hat{P}_i^4]^\mathsf{T}$. 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 et al., 2018a,b, 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 distortions 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 the jackknife standard error, which we show is also consistent in our setting even when many covariates are used. Therefore, implementation is straightforward and is available in any statistical computing software. 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.

REFERENCES

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

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1–19.

Abadie, A. and M. D. Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*, 465–503.

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

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*(4), 962–972.

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.

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 (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association 113*(522), 767–779.

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 (2010). Robust data-driven inference for density-weighted average derivatives. *Journal of the American Statistical Association 105*(491), 1070–1083.

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., R. K. Crump, and M. Jansson (2014a). Bootstrapping density-weighted average derivatives. *Econometric Theory 30*(6), 1135–1164.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2014b). Small bandwidth asymptotics for density-weighted average derivatives. *Econometric Theory 30*(1), 176–200.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2018). Large sample properties of partitioning-based estimators. arXiv:1804.04916.

Cattaneo, M. D. and M. Jansson (2018). Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency. *Econometrica 86*(3), 955–995.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory 34*(2), 277–301.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018b). 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*. New York: Wiley.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume VI*, pp. 5549–5632. New York: Elsevier Science B.V.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J. M. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. R. Robins (2018). Locally robust semiparametric estimation. arXiv:1608.00033.

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.

Fernandez-Val, I. and M. Weidner (2018). Fixed effects estimation of large-*t* panel data models. *Annual Review of Economics 10*.

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.

Ichimura, H. and P. E. Todd (2007). Implementing nonparametric and semiparametric estimators. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume VIB*, 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.

Mammen, E. (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 R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics, Volume IV*, pp. 2111–2245. New York: Elsevier Science B.V.

Newey, W. K. and J. M. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv:1801.09138.

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* (2 ed.). Cambridge, MA: MIT Press.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources 50*(2), 420–445.
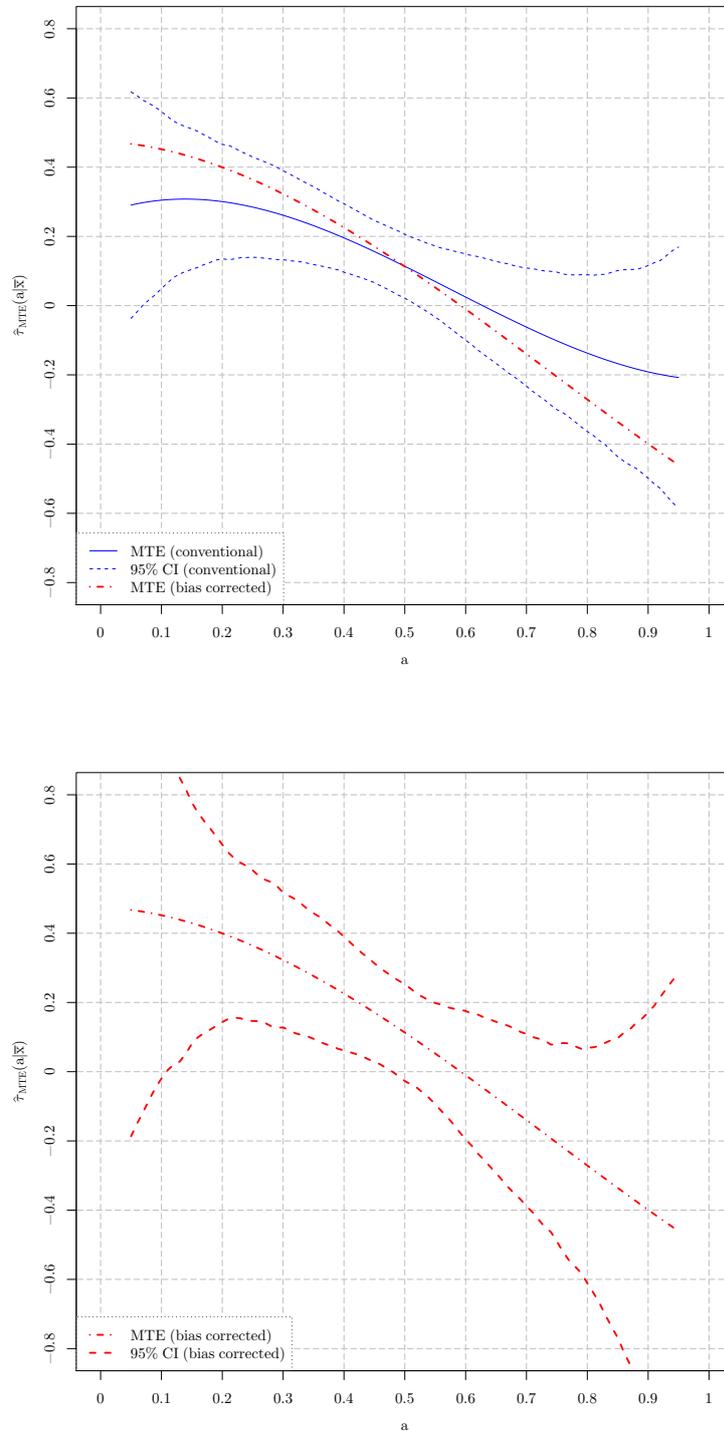
FIGURE 1

Estimated marginal treatment effects. The marginal treatment effect, $\hat{\tau}_{\texttt{MTE}}(a|\bar{\mathbf{X}})$, is evaluated at mean value of the covariates. Bootstrap is used to construct the confidence interval, with 500 repetitions. *Top*: 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). *Bottom:* Bias-corrected MTE, together with 95% confidence interval, taking into account the effect of bias correction.

TABLE 1

Simulation: marginal treatment effects. 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.

k	k/n	$k/\sqrt{n}$	Conventional					Bias-Corrected				
			bias	sd	$\sqrt{mse}$	coverage	length	bias	sd	$\sqrt{mse}$	coverage	length
							Panel (a) $n = 1000$					
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
							Panel (b) $n = 2000$					
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

[1] k: number of instruments used for propensity score estimation;
[2] bias: empirical bias (scaled by $\sqrt{n}$);
[3] sd: empirical standard deviation (scaled by $\sqrt{n}$);
[4] $\sqrt{mse}$: empirical root-MSE (scaled by $\sqrt{n}$);
[5] 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);
[6] length: the average confidence interval length (scaled by $\sqrt{n}$).