An Averaging Estimator for Two Step M Estimation in Semiparametric Models

Ruoyao Shi*

September, 2022

*Department of Economics, University of California Riverside, ruoyao.shi@ucr.edu. The comments from the editor Peter C. B. Phillips, the associate editor Patrik Guggenberger and two anonymous referees were vastly helpful in improving this paper. The author also thanks Colin Cameron, Xu Cheng, Denis Chetverikov, Yanqin Fan, Jinyong Hahn, Bo Honoré, Toru Kitagawa, Zhipeng Liao, Hyungsik Roger Moon, Whitney Newey, Geert Ridder, Aman Ullah, Haiqing Xu and the participants at various seminars and conferences for helpful comments. This project is generously supported by UC Riverside Regents' Faculty Fellowship 2019-2020. Zhuozhen Zhao provided great research assistance. All remaining errors are the author's.

Running head: Two step SP averaging estimator
Corresponding author: Ruoyao Shi (ruoyao.shi@ucr.edu)

Abstract

In a two step extremum estimation (M estimation) framework with a finite dimensional parameter of interest and a potentially infinite dimensional first step nuisance parameter, this paper proposes an averaging estimator that combines a semiparametric estimator based on nonparametric first step and a parametric estimator which imposes parametric restrictions on the first step. The averaging weight is an easy-to-compute sample analog of an infeasible optimal weight that minimizes the asymptotic quadratic risk. Under Stein-type conditions, the asymptotic lower bound of the truncated quadratic risk difference between the averaging estimator and the semiparametric estimator is strictly less than zero for a class of data generating processes (DGPs) that includes both correct specification and varied degrees of misspecification of the parametric restrictions, and the asymptotic upper bound is weakly less than zero. The averaging estimator, along with an easy-to-implement inference method, is demonstrated in an example.

Keywords: two step M estimation, semiparametric model, averaging estimator, uniform dominance, asymptotic quadratic risk
JEL Codes: C13, C14, C51, C52

1 Introduction

Semiparametric models, consisting of a parametric component and a nonparametric component, have gained popularity in economics. Being approximations of complex economic activities, they harmoniously deliver two advantages at the same time: parsimonious modeling of parameters of interest and robustness against misspecification of arbitrary parametric restrictions on activities that are not central for the research question at hand. One disadvantage of associated semiparametric estimators, however, is that they are typically less efficient than their parametric counterparts which result from imposing certain parametric restrictions on the nonparametric components of semiparametric models.[1] This efficiency defect of semiparametric estimators often renders relatively imprecise estimates and low test power, especially when the parametric restrictions are correct or only mildly misspecified.

Recognizing such tension between robustness and efficiency, researchers have utilized various specification tests to choose between semiparametric and parametric estimators in practice. Neither parametric estimators nor the resulting pre-test estimators, however, are robust to misspecification of the parametric restrictions, since whether they are more accurate than the semiparametric estimators depends on the unknown degree of misspecification.

This paper aims to solve this tension between robustness and efficiency in semiparametric models by developing an estimator whose improvement on the accuracy over semiparametric estimators (used as benchmark) is robust against varied degrees of misspecification of the parametric restrictions. First, this paper proposes an averaging estimator that is a simple weighted average between the semiparametric estimator and the parametric estimator with a data-driven weight. Second, under mild Stein-type conditions, the proposed averaging estimator is proven to have (weakly) smaller asymptotic quadratic risks – a general class of measures of accuracy that includes mean squared error (MSE) as a special case – than the semiparametric benchmark regardless of whether the parametric restrictions are correct or misspecified, and regardless of the degree of misspecification. Third, an inference method that is valid regardless of the unknown degree of misspecification is recommended.

Let β denote the unknown parameter of interest, and let $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ denote the semiparametric and the parametric estimators, respectively. The averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ takes the form

$$\hat{\beta}_{n,\hat{w}_n} \equiv (1-\hat{w}_n)\hat{\beta}_{n,SP} + \hat{w}_n\hat{\beta}_{n,P}, \tag{1.1}$$

where n is the sample size and $\hat{w}_n$ is a data-driven averaging weight elaborated in (2.2) below. Intuitively, the weight quantifies the asymptotic efficiency gain achieved by imposing the parametric restrictions and the possible asymptotic misspecification bias incurred by deviating from the robust semiparametric benchmark. It then balances the two to reduce asymptotic quadratic risks compared to the semiparametric estimator.

This paper employs a *uniform* asymptotic theory to approximate the upper and the lower bounds of the finite-sample truncated quadratic risk difference between the averaging estimator and the semiparametric estimator over a large class of DGPs.[2] Extending the subsequence argument developed in Cheng, Liao, and Shi (2019) for generalized method of moments (GMM) estimators, this paper shows that the sufficient conditions for the lower bound to be strictly less than zero and for the upper bound to be weakly less than zero are mild. Since this class of DGPs include those under which the parametric restrictions are correctly specified, mildly misspecified and severely misspecified, the uniform dominance result here asserts that the

[1] This paper will use the terms "parametric estimator" and "parametric restrictions" loosely. They do not necessarily mean that the data distribution is fully parametric, but only mean that the nonparametric argument in the estimation objective function belongs to a finite dimensional subspace of certain infinite dimensional function space, as described in (3.5) below.

[2] The loss function and the truncated loss function are defined in (2.1) and (3.10), respectively.

averaging estimator achieves improvement in accuracy over the semiparametric estimator in a way that is robust against varied degrees of misspecification. Unlike Cheng et al. (2019) which focuses on one step GMM estimators, this paper considers two step M estimation framework for semiparametric models as it encompasses maximum likelihood estimator (MLE), GMM, many kernel-based and sieve estimators, etc. as special cases, as well as regular one step M estimators.

The semiparametric models considered in this paper are flexible enough to include many popular models as special examples – such as single-index models (Ahn, Ichimura, and Powell, 1996), transformation models (Han, 1987; Sherman, 1993), censored and truncated regression models (Powell, 1986), control function approaches (Blundell and Powell, 2003, 2004), nonlinear panel data models (Honoré, 1992), and dynamic discrete choice models (Hotz and Miller, 1993; Keane and Wolpin, 1997; Buchholz, Shum, and Xu, 2021), among others.

The proposed averaging estimator is demonstrated using a carefully curated partially linear model example. A point worth emphasizing here is that although the *estimation error* of the nonparametric component does not affect the asymptotic properties of the parametric component estimator in partially linear models (Robinson, 1988), the *presence* of the nonparametric component and how it is *modeled* still generally inflicts critical impacts on the latter. This point will become clearer in Section 4.

This paper has a few obvious limitations. First, the uniform asymptotic dominance result in this paper does not guarantee that the averaging estimator outperforms the semiparametric benchmark in finite samples, even though the uniform asymptotic analysis employed here provides better approximation of the estimators' finite sample properties than the usual pointwise asymptotic framework. Second, inference based on the proposed averaging estimator, like most cases (if not all) of post-averaging inference, is more challenging than that based on standard estimators. The two step method proposed by Claeskens and Hjort (2008) is used to construct an asymptotically valid confidence interval in this paper (also see, e.g., Kitagawa and Muris, 2016, for its application), but its coverage probability can be conservative. Third, this paper focuses on averaging between one semiparametric estimator and one parametric estimator, excluding estimators that average the semiparametric estimator with more than one parametric estimator and potentially outperform the one proposed in this paper. These limitations all point out important directions for future research.

Related literature. This paper belongs to the growing literature on frequentist shrinkage and model averaging estimators, which are weighted averages of other estimators.[3] Shrinkage estimators date back to the James-Stein estimator in Gaussian models (James and Stein, 1961), and are comprehensively reviewed by Fourdrinier, Strawderman, and Wells (2018). Recent years have seen development of frequentist model averaging estimators in many contexts. Hjort and Claeskens (2003) and Hansen (2016) consider likelihood-based estimators in parametric models. In least square regression models, various model averaging estimators are developed and their properties are carefully examined by Judge and Mittelhammer (2004); Mittelhammer and Judge (2005); Hansen (2007); Wan, Zhang, and Zou (2010); Hansen and Racine (2012); Hansen (2014); Liu (2015) and Hansen (2017), just to name a few. Lu and Su (2015) study quantile regression models. For semiparametric models, Judge and Mittelhammer (2007); DiTraglia (2016) consider averaging GMM estimators, and Kitagawa and Muris (2016) analyze averaging semiparametric estimators of the treatment effects on the treated (ATT) based on different parametric propensity score models. Averaging estimators in nonparametric models are also discussed, for example, by Fan and Ullah (1999); Yang (2001, 2003); Wasserman (2006) and Peng and Yang (2021). Magnus, Powell, and Prüfer (2010) and Fessler and Kasy (2019), among others, investigate Bayesian model averaging estimators as well. Claeskens and Hjort (2008) provide an excellent review of both frequentist and Bayesian model averaging estimators. My paper differs from this

[3]Such names as combined or ensemble estimators are also used by different authors to refer to weighted averages of other estimators with different goals and approaches.

literature in the following ways. First,the two step semiparametric M estimation framework in this paper nests many familiar estimators (one step or two step) in semiparametric (and parametric) models as special cases. Second, in contrast to the literature on nonparametric models that deals with unknown functions and averages among growing number of estimators, this paper focuses on finite dimensional parameters in semiparametric (and parametric) models and averages between two estimators. The asymptotic theories of the two differ substantially. Third, the averaging weight, when specialized to corresponding cases, differs from those in the aforementioned papers. Fourth, the proposed averaging estimator is shown to dominate the semiparametric benchmark under a uniform asymptotic approach, instead of the pointwise local asymptotic approach (Le Cam, 1972; Van der Vaart, 2000, Chapter 7) often taken in the literature. Finally, when a certain weighting matrix is used in the loss function, one of the sufficient conditions for the uniform dominance of the averaging estimator is of the common Stein type. This condition is stronger than the Stein-type conditions for some shrinkage estimators in the literature and weaker than others.[4]

This paper is particularly related to Cheng et al. (2019), but it generalizes their uniform asymptotic approach and the subsequence technique from one step GMM estimators in moment condition models to two step M estimators in more general semiparametric models.[5] Moreover, the restricted estimator considered in Cheng et al. (2019) is asymptotically efficient, but this paper allows the restricted (parametric) estimator to be away from the efficiency bound. This relaxation is useful in practice since in complex semiparametric models, the efficient estimators under the parametric restrictions may be difficult to implement or may have certain undesirable features, and the widely used ones may fall short of the efficiency bound (e.g., the Heckit estimator in sample selection models).

The uniform asymptotic analysis in this paper is premised upon high-level asymptotic distributions of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$, which can be justified under various primitive conditions in different models, as shown in numerous previous studies on the asymptotic properties of specific and general M estimators – e.g., Lee (1982); Gallant and Nychka (1987); Ahn and Powell (1993); Newey and Powell (1993); Andrews (1994); Newey (1994); Newey and McFadden (1994); Powell (1994); Pakes and Olley (1995); Powell (2001); Bickel and Ritov (2003); Chen, Linton, and Van Keilegom (2003); Hirano, Imbens, and Ridder (2003); Firpo (2007); Newey (2009); Ichimura and Lee (2010); Ackerberg, Chen, and Hahn (2012); Ackerberg, Chen, Hahn, and Liao (2014) and Ichimura and Newey (2017) – among others.

Averaging estimators can be regarded as a smoothed generalization of pre-test estimators (or model selection estimators), since the latter restrict the averaging weights to be either zero or one depending on the result of certain specification tests or criteria. For models involving infinite dimensional components, many authors propose various specification tests, including Bierens (1990)using sieve estimators, Robinson (1989) using kernel estimators, and many others. Model selection estimators based on focused information criterion (FIC) in semiparametric models are considered, for example, by Hjort and Claeskens (2006). Pre-test estimators typically perform better than the unrestricted benchmark for certain degrees of misspecification of the restrictions and worse for the others. Moreover, the literature has documented that in many settings, the maximal scaled quadratic risks of pre-test estimators based on consistent tests grow unbounded as sample sizes increase, despite promising properties suggested by pointwise asymptotic analysis. A well-cited example is the Hodges' estimator (e.g., Van der Vaart, 2000, Example 8.1), among others (Yang, 2005; Leeb and Pötscher, 2005, 2008; Hansen, 2016; Cheng et al., 2019, etc.). In contrast, the uniform asymptotic approach of this paper better approximates the finite sample properties of the averaging estimator, so the resulting averaging estimator has (weakly) smaller asymptotic quadratic risks than the semiparametric benchmark uniformly over the degree of misspecification and avoids the common pitfalls of pre-test estimators.

[4]This sufficient condition, like for James-Stein estimator and many other shrinkage estimators, requires the dimension of the parameters of interest to be larger than a certain number. See the discussion after Theorem 1 in Section 3 for details.

[5]Cheng et al. (2019) is in turn based on the uniform inference analysis in Andrews, Cheng, and Guggenberger (2011).

This paper is related to but differs from the following strands of literature as well. First, doubly robust estimators in statistics (e.g., Scharfstein, Rotnitzky, and Robins, 1999; Bang and Robins, 2005; Rubin and van der Laan, 2008; Cao, Tsiatis, and Davidian, 2009; Tsiatis, Davidian, and Cao, 2011) are robust against misspecification, but they typically require that some components of the model are correctly specified, while the averaging estimator in this paper exhibits improved risk regardless of the degree of misspecification. Second, recent development in locally robust estimators in semiparametric models (e.g., Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018; Chernozhukov, Escanciano, Ichimura, Newey, and Robins, 2022) removes impacts of the nuisance function estimation bias (brought by regularization of machine learning methods) on the influence function of the parameter of interest by orthogonalization (Neyman, 1959). The approach here is still useful, since how the nuisance function is *modeled* affects both variance and bias even when it is *known* and needs no estimation. Third, among the literature on sensitivity analysis (e.g., Rosenbaum and Rubin, 1983; Leamer, 1985; Imbens, 2003; Altonji, Elder, and Taber, 2005; Andrews, Gentzkow, and Shapiro, 2017; Mukhin, 2018; Oster, 2019), Bonhomme and Weidner (2018) and Armstrong and Kolesár (2021) are the closest to this paper. They take a restricted model as benchmark and study the sensitivity of the results with respect to possible local misspecification that deviates from it. This paper takes an opposite perspective by positing a robust unrestricted semiparametric model as benchmark and pursuing uniform quadratic risk improvement with the help of added parametric restrictions.

Plan of the paper. The rest of this paper is organized as follows. Section 2 prescribes how to compute the proposed averaging estimator in practice. Section 3 states and proves the main uniform dominance result of the paper along with its conditions and an inference method. Section 4 conducts Monte Carlo (MC) experiments using a partially linear model example to investigate the finite sample performance of the proposed averaging estimator. Section 5 concludes. Appendix A gives the proofs of the results in Section 3. Appendix B provides the proofs of the intermediate lemmas in Appendix A. Appendix C presents an alternative method of computing the averaging weight. Appendix D details additional theoretical and Monte Carlo results for the example in Section 4. Appendix E discusses the justification for the high-level Condition 2. Appendices B to E are available online in the supplementary material associated with this article.

2 Averaging Weight

This section explains how to compute the averaging estimator in practice. Rigorous conditions and the formal uniform asymptotic dominance result will be provided in Section 3.

One is interested in the estimation of a finite dimensional vector of parameters $\beta \in \mathcal{B}$, where $\mathcal{B} \subset \mathbb{R}^k$ is compact. Let $\mathcal{F}$ denote the set of DGPs, and let F denote one DGP from $\mathcal{F}$. For any estimator $\hat{\beta}_n$ of the parameter β and a chosen symmetric positive semi-definite weighting matrix Υ,[6] define the loss function to be the following quadratic form[7]

$$\ell(\hat{\beta}_n, \beta) \equiv n(\hat{\beta}_n - \beta)'\Upsilon(\hat{\beta}_n - \beta). \tag{2.1}$$

Here the weighting matrix Υ is chosen by the researcher and reflects how much the researcher values the estimation accuracy of each coordinate of β.[8] If the researcher treats every coordinate equally, then they

[6] Υ can be assumed to be symmetric without loss of generality, because for any asymmetric $\tilde{\Upsilon}$ there exists a symmetric Υ that gives rise to the same loss function.

[7] Hansen (2016) argues that the choice of loss function affects asymptotic performance of estimators only via its local quadratic approximation, so considering a quadratic loss function is not as restrictive as it may appear. To be precise, the loss function used in the asymptotic theory of this paper is a truncated version of (2.1), which is defined in (3.10) below.

[8] With certain choice of Υ, the main theorem of this paper (Theorem 1 below) requires $k \geq 4$ (see the discussion that immediately follows Theorem 1).

may choose $\Upsilon = I_k$ (the $k \times k$ identity matrix). If the researcher focuses on the prediction error in a linear regression model, then they may choose $\Upsilon = \mathbb{E}_F(X_i X_i')$, where $\mathbb{E}_F(\cdot)$ denotes the expectation operator under DGP F and X_i denotes the regressors. If the researcher focuses on only a subvector of β, then they may choose Υ to be a diagonal matrix with diagonal entries associated with the subvector being one and other diagonal entries being zero. This last example shares the same spirit with the FIC model averaging (Zhang and Liang, 2011), but the weighting matrix Υ here affords more flexibility. Note that both the loss function and the averaging weight (to be introduced later) depend on Υ, but such dependence is suppressed for notational simplicity.

Given the loss function in (2.1), the semiparametric estimator $\hat{\beta}_{n,SP}$ is preferred in terms of robustness since it is consistent whether the parametric restrictions hold or not. The parametric estimator $\hat{\beta}_{n,P}$ is consistent only if those restrictions are sufficiently close to holding, and if they do, $\hat{\beta}_{n,P}$ will typically be asymptotically more efficient than $\hat{\beta}_{n,SP}$. As a result, the potentially more efficient $\hat{\beta}_{n,P}$ sometimes has improved asymptotic quadratic risk over the robust $\hat{\beta}_{n,SP}$ but sometimes does not. The optimal robustness-efficiency trade-off (i.e., bias-variance trade-off) depends on the degree of misspecification of the parametric restrictions, a measurement unknown to the researcher.

The main message of this paper is that with the proposed averaging weight, the averaging estimator of the form in (1.1) *always* has no larger asymptotic risk than the robust estimator $\hat{\beta}_{n,SP}$ *regardless* of whether the parametric restrictions hold or not and *regardless* of the degree of misspecification.

Under DGP F, let $V_{F,SP}$ and $V_{F,P}$ be the asymptotic variance-covariance matrices of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$, respectively, and let C_F be their asymptotic covariance matrix. Let $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$ be the consistent estimators. Then the data-driven averaging weight is

$$\hat{w}_n \equiv \frac{\text{tr}[\Upsilon(\widehat{V}_{n,SP} - \widehat{C}_n)]}{\text{tr}[\Upsilon(\widehat{V}_{n,SP} + \widehat{V}_{n,P} - 2\widehat{C}_n)] + n(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})'\Upsilon(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})}, \tag{2.2}$$

where $\text{tr}(\cdot)$ indicates the trace of a square matrix.[9] This weight falls in the interval $[0, 1]$ with probability one, and the reason is detailed in Appendix C.[10]

Remark 1. *If $\hat{\beta}_{n,P}$ is an asymptotically efficient estimator under the parametric restrictions, then $C_F = V_{F,P}$. In this case, the averaging weight can simplify to*

$$\hat{w}_n \equiv \frac{tr[\Upsilon(\widehat{V}_{n,SP} - \widehat{V}_{n,P})]}{tr[\Upsilon(\widehat{V}_{n,SP} - \widehat{V}_{n,P})] + n(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})'\Upsilon(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})}, \tag{2.3}$$

which resembles the GMM averaging weight proposed by Cheng et al. (2019).

It is easier to see the intuition of the averaging weight from (2.3). If the asymptotic efficiency gain of imposing the first step parametric restrictions, represented by $\text{tr}[\Upsilon(\widehat{V}_{F,SP} - \widehat{V}_{F,P})]$, is large, then the averaging estimator ought to allocate more weight to $\hat{\beta}_{n,P}$. If, on the other hand, the asymptotic bias of $\hat{\beta}_{n,P}$ resulting from misspecification of the restrictions, represented by $\hat{\beta}_{n,P} - \hat{\beta}_{n,SP}$ (since $\hat{\beta}_{n,SP}$ is always consistent), is large, then the averaging estimator should assign less weight to $\hat{\beta}_{n,P}$. The proposed weight in (2.3) operationalizes such intuition by striking a balance between robustness and efficiency.

The weight in (2.2) generalizes (2.3) by allowing for averaging even when $\hat{\beta}_{n,P}$ is not asymptotically efficient. This generalization is especially important for semiparametric models, because asymptotically efficient

[9] Note that in (2.2), $\widehat{C}_n$ is in general an asymmetric matrix, i.e., $\widehat{C}_n \neq \widehat{C}_n'$, but $\Upsilon\widehat{C}_n$ and $\Upsilon\widehat{C}_n'$ have the same trace due to the symmetry of Υ and properties of the trace operator. The same goes for C_F and C_F'.

[10] In finite samples, however, it is possible that $\hat{w}_n$ falls outside the interval $[0, 1]$. One could add a restriction that enforces $\hat{w}_n \in [0, 1]$, and this can be justified by minor changes (not detailed in this paper) in the proofs of the theoretical results in Section 3. The author thanks an anonymous referee for pointing this out.

estimators do not always exist in these models, and might be difficult to compute or possess undesirable finite sample properties when they do. A salient example is the sample selection model under the joint normality restriction, where the Heckman (1979) two step Heckit estimator is asymptotically inefficient but more widely used than the efficient MLE, for a variety of reasons (see the discussion in Heckman, 1976; Wales and Woodland, 1980; Nelson, 1984).

Bootstrapping asymptotic variance-covariance matrices.[11] The key to the construction of the averaging weight $\hat{w}_n$, as (2.2) implies, is the consistent variance-covariance matrix estimators $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$. They can be computed by bootstrapping the asymptotic variance-covariance matrix of $\left(\hat{\beta}'_{n,SP}, \hat{\beta}'_{n,P}\right)'$.

Because the consistency of the bootstrap distribution does not guarantee the consistency of the bootstrap second moment (Hahn and Liao, 2021), one needs to use one of the consistent bootstrap variance-covariance estimators proposed in the literature instead of the simple bootstrap second moment. Among them, the following truncation method proposed by Shao (1992) and adapted to this paper is both general and easy to implement.

(1) Let $\hat{\beta} \equiv \left(\hat{\beta}'_{n,SP}, \hat{\beta}'_{n,P}\right)'$ and let $\hat{\beta}_j$ $(j = 1, \ldots, 2k)$ denote its jth coordinate. For a larger number B, randomly draw B bootstrap samples of size n and compute the bootstrap estimate $\hat{\beta}^b$ $(b = 1, \ldots, B)$ for each sample.

(2) For fixed positive constants ν and small c_0, define $T_j \equiv \max\{\nu|\hat{\beta}_j|, c_0\}$ for each coordinate.[12] For all b and all j, define

$$\Delta_j^b \equiv \begin{cases} T_j, & if\ \hat{\beta}_j^b - \hat{\beta}_j > T_j; \\ \hat{\beta}_j^b - \hat{\beta}_j, & if\ |\hat{\beta}_j^b - \hat{\beta}_j| \leq T_j; \\ -T_j, & if\ \hat{\beta}_j^b - \hat{\beta}_j < -T_j; \end{cases}$$

and $\Delta^b \equiv \left(\Delta_1^b, \ldots, \Delta_{2k}^b\right)'$.

(3) Compute the $2k \times 2k$ matrix $\widehat{V}_n \equiv \frac{n}{B}\sum_{b=1}^{B}(\Delta^b - \bar{\Delta})(\Delta^b - \bar{\Delta})'$, where $\bar{\Delta} \equiv B^{-1}\sum_{b=1}^{B}\Delta^b$. Then $\widehat{V}_{n,SP}$ is the upper left $k \times k$ block of $\widehat{V}_n$, $\widehat{V}_{n,P}$ is the lower right $k \times k$ block of $\widehat{V}_n$, and $\widehat{C}_n$ is the upper right $k \times k$ block of $\widehat{V}_n$.

These bootstrapped asymptotic variance-covariance matrix estimates can then be plugged into (2.2) to compute the averaging weight. Appendix C provides an alternative method of computing the asymptotic variance-covariance matrices via the influence functions of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$, and Section 4 will show that the finite sample performance of the averaging estimator using the influence functions and that using the bootstrapping method are almost identical to each other.

3 Theoretical Results

This section proves and provides the conditions for the uniform dominance result of the averaging estimator. An inference method is also suggested.

Suppose β_F, the true parameter value under DGP F, is identified as the unique minimizer (assume it exists) of some objective function $Q_F(\beta, h_F)$; in other words, the parameter of interest is

$$\beta_F \equiv \arg\min_{\beta \in \mathcal{B}} Q_F(\beta, h_F), \tag{3.1}$$

[11] The author thanks an anonymous referee for suggesting providing a bootstrap method for computing the averaging weight.

[12] The validity of Shao (1992)'s method does not rely on any specific values of ν or c_0. In his paper, $\nu = 1$ and $c_0 = 0.05$ were used in the simulation study. These values will also be used in the Monte Carlo experiments in Section 4 of this paper.

where the objective function $Q_F(\beta, h)$ depends on some potentially infinite dimensional nuisance parameter h. Since the objective function Q_F has h as an argument, the presence of h and how it is modeled generally affect the asymptotic properties of β estimators through Q_F, even in the absence of estimation error of h.[13] Under DGP F, the true nuisance parameter value h_F is identified as the unique minimizer (assume it exists) of another objective function $R_F(h)$; that is,

$$h_F \equiv \arg\min_{h \in \mathcal{H}} R_F(h), \tag{3.2}$$

where $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is some complete, separable space of square integrable functions of data Z.

A general class of two step M estimators $\hat{\beta}_n$ is as follows,

$$\hat{\beta}_n \equiv \arg\min_{\beta \in \mathcal{B}} \hat{Q}_n(\beta, \hat{h}_n), \tag{3.3}$$

where $\hat{Q}_n(\beta, \hat{h}_n)$ is some empirical objective function of β which depends on the sample $\{Z_i\}_{i=1}^n$ and $\hat{h}_n$, a first step estimator of the unknown nuisance parameter h. Throughout this paper, the dependence of the empirical objective functions on the sample $\{Z_i\}_{i=1}^n$ is suppressed for notational simplicity.

As alluded before, this paper considers averaging two common two step M estimators of β, the semiparametric estimator $\hat{\beta}_{n,SP}$ and the parametric estimator $\hat{\beta}_{n,P}$. The two differ in how h is modeled and estimated in the first step. The semiparametric estimator $\hat{\beta}_{n,SP}$ does not impose specific functional form restrictions on h, so $\hat{h}_n$ results from common nonparametric estimation procedures. For example, suppose $\hat{h}_n$ is obtained from a first step sieve M estimation procedure as follows,

$$\hat{h}_n \equiv \arg\min_{h \in \mathcal{H}_n} \hat{R}_n(h), \tag{3.4}$$

where $\hat{R}_n(h)$ is some empirical objective function, and $\mathcal{H}_n$ are subspaces of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ that become dense as $n \to \infty$. Then the first step of the corresponding $\hat{\beta}_{n,SP}$ is (3.4) and the second step is (3.3).

On the other hand, the parametric estimator $\hat{\beta}_{n,P}$ arises because economic hypotheses often suggest a certain parametric form of h, or one may want to limit the dimension of h to improve the efficiency. In these cases, one will model h with a finite dimensional subspace of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, denoted as $\mathcal{H}_g$, where g is a function that is known up to a finite dimensional vector of unknown parameters γ. Formally, let $\Gamma \subset \mathbb{R}^t$ be a compact subset of the t-dimensional Euclidean space, then

$$\mathcal{H}_g \equiv \{h(\cdot) : \exists \text{ some } \gamma \in \Gamma \text{ such that } h(\cdot) \in \mathcal{H} \text{ and } h(\cdot) = g_\gamma(\cdot) \equiv g(\cdot; \gamma)\}. \tag{3.5}$$

Let

$$\hat{\gamma}_n \equiv \arg\min_{\gamma \in \Gamma} \hat{R}_n(g_\gamma), \tag{3.6}$$

and let the restricted nuisance parameter estimate be written as $\hat{h}_n = g_{\hat{\gamma}_n}$. Then the first step of $\hat{\beta}_{n,P}$ is (3.6) and the second step is (3.3).

For every DGP $F \in \mathcal{F}$ and under the parametric restrictions, define the first step pseudo-true parameter vector γ_F as the unique minimizer (assume it exists) of the following problem

$$\gamma_F \equiv \arg\min_{\gamma \in \Gamma} R_F(g_\gamma), \tag{3.7}$$

[13] Typically, the influence function of the estimator $\hat{\beta}_n$ depends on the first and second derivatives of Q_F, which in turn both depend on h generally (see, e.g. Newey, 1994; Ichimura and Lee, 2010; Ackerberg et al., 2014; Ichimura and Newey, 2017).

where the first step objective function $R_F(\cdot)$ is the same as in (3.2) and the first step nuisance function subspace $\mathcal{H}_g$ is defined in (3.5). Also define the second step pseudo-true parameter $\beta_{F,P}$ as the unique minimizer (assume it exists) of the following problem

$$\beta_{F,P} \equiv \arg\min_{\beta \in \mathcal{B}} Q_F(\beta, g_{\gamma_F}), \tag{3.8}$$

where $Q_F(\cdot,\cdot)$ is the same as in (3.1). In general, the nuisance function g_{γ_F} induced by the pseudo-true parameter γ_F is different from the true nuisance function h_F identified in (3.2). In consequence, $\beta_{F,P}$ in general will be different from β_F, the true parameter of interest identified in (3.1). Let δ_F denote the bias caused by imposing the parametric restrictions; that is

$$\delta_F \equiv \beta_{F,P} - \beta_F. \tag{3.9}$$

The key to the uniform dominance is to determine the sign of the asymptotic risk difference between the averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ and the semiparametric estimator $\hat{\beta}_{n,SP}$ under DGPs with varied degrees of misspecification. This paper utilizes the uniform asymptotic approach and the subsequence technique in Cheng et al. (2019), instead of Pitman sequences, which is frequently used when analyzing the pointwise local asymptotic properties of estimators. Lower (infimum) and upper (supremum) bounds of the risk differences between $\hat{\beta}_{n,\hat{w}_n}$ and $\hat{\beta}_{n,SP}$ for all DGPs within a set $\mathcal{F}$ satisfying certain regularity conditions are considered before rendering the sample size to infinity.

To formally state the dominance result, some notation is needed. For any estimator $\hat{\beta}_n$ of β and an arbitrary real number ζ, define the truncated loss function

$$\ell_\zeta(\hat{\beta}_n, \beta) \equiv \min\{\ell(\hat{\beta}_n, \beta), \zeta\}, \tag{3.10}$$

where $\ell(\hat{\beta}_n, \beta)$ is the quadratic loss function defined in (2.1). Compared to the loss function in (2.1), the truncation does not restrict the applicability of the main result much as ζ could be arbitrarily large. The bounds of the truncated risk differences for finite sample size n are defined as:

$$\underline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta) \equiv \inf_{F \in \mathcal{F}} \mathbb{E}_F[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_F) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_F)], \tag{3.11}$$

$$\overline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta) \equiv \sup_{F \in \mathcal{F}} \mathbb{E}_F[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_F) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_F)]. \tag{3.12}$$

Then, define the following limits of the finite sample bounds:

$$Asy\underline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \equiv \lim_{\zeta \to \infty} \liminf_{n \to \infty} \underline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta), \tag{3.13}$$

$$Asy\overline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \equiv \lim_{\zeta \to \infty} \limsup_{n \to \infty} \overline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta). \tag{3.14}$$

The key difference between these bounds and the asymptotic risks that utilize Pitman sequences in pointwise local analysis is that the truncated risk differences in (3.11) and (3.12) are extrema over the entire DGP set $\mathcal{F}$ for each finite sample size n, before n is sent to infinity to obtain the asymptotic bounds in (3.13) and (3.14). The finite sample extrema may occur at different Pitman sequences for different n, allowing the asymptotic bounds to be approached not along a single Pitman sequence.

The averaging estimator is said to dominate the semiparametric estimator in terms of asymptotic truncated risk uniformly over $\mathcal{F}$ if

$$Asy\underline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) < 0, \tag{3.15}$$

and

$$Asy\overline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \leq 0. \tag{3.16}$$

(3.15) and (3.16) will be shown to hold in Theorem 1 under the following conditions and intermediate results.

Condition 1. *Recall δ_F defined in* (3.9)*. Suppose $\mathcal{F}$ is such that the following holds.*
(i) $\delta_F = 0$ only if $h_F = g_{\gamma_F}$ for some $\gamma_F \in \mathbb{R}^t$;
(ii) $0_{k\times 1} \in int(\Delta_\delta)$, where $\Delta_\delta \equiv \{\delta_F \colon F \in \mathcal{F}\}$.

Condition 1(i) is a simple requirement that if the parametric restrictions on the nuisance function h are misspecified, then the pseudo-true parameter value $\beta_{F,P}$ will differ from the true value β_F, which rules out the uninteresting special case that β_F may be consistently estimable even with severely misspecified parametric restrictions. As a result, the degree of misspecification can be indexed by δ_F, the bias introduced by imposing the parametric restrictions. Condition 1(ii) says that the parametric restrictions may be misspecified with varied degrees, including the correct specification case. Condition 1 does not impose any stringent restrictions on the semiparametric models.

Use the following notation for the nuisance parameter vector that characterizes the joint asymptotic distributions of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ under DGP F,

$$\bar{S}(F) \equiv (\text{vech}(V_{F,SP})', \text{vech}(V_{F,P})', \text{vec}(C_F)')', \text{ and } S(F) \equiv \left(\delta_F', \bar{S}(F)'\right)', \tag{3.17}$$

where δ_F is defined in (3.9), and vech$(\cdot)$ and vec$(\cdot)$ are vectorization of distinct elements of a matrix. Let

$$\mathcal{S} \equiv \{S(F) \colon F \in \mathcal{F}\}. \tag{3.18}$$

A sequence of DGPs $\{F_n\}_{n=1}^\infty$ is said to be *correctly specified* if $n^{1/2}\delta_{F_n} \to 0$, *locally / mildly misspecified* if $n^{1/2}\delta_{F_n} \to d \in (0, \infty)$, and *severely misspecified* if $n^{1/2}\delta_{F_n} \to \infty$.

Condition 2. *For any sequence of DGPs $\{F_n\}_{n=1}^\infty$ such that $\bar{S}(F_n) \to \bar{S}(F)$ for some $F \in \mathcal{F}$ and $n^{1/2}\delta_{F_n} \to d \in \mathbb{R}_\infty^k$, suppose the estimators $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ satisfy the following conditions.*
(i) If $\|d\| < \infty$, then

$$\begin{bmatrix} n^{1/2}(\hat{\beta}_{n,SP} - \beta_{F_n}) \\ n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n}) \end{bmatrix} \xrightarrow{d.} \begin{bmatrix} \xi_{F,SP} \\ \xi_{F,P} + d \end{bmatrix}, \tag{3.19}$$

where $\tilde{\xi}_F \sim \mathcal{N}(0_{2k\times 1}, \tilde{V}_F)$, with $\tilde{\xi}_F \equiv (\xi_{F,SP}', \xi_{F,P}')'$, $V_{F,SP} \geq V_{F,P}$ and

$$\tilde{V}_F \equiv \begin{bmatrix} V_{F,SP} & C_F \\ C_F & V_{F,P} \end{bmatrix}.$$

(ii) If $\|d\| = \infty$, then $n^{1/2}(\hat{\beta}_{n,SP} - \beta_{F_n}) \xrightarrow{d.} \xi_{F,SP}$ and $\|n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n})\| \xrightarrow{p.} \infty$.

Condition 2(i) requires that both $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are locally regular estimators (Ichimura and Newey, 2017, Definition 1), which means that $n^{1/2}((\hat{\beta}_{n,SP}', \hat{\beta}_{n,P}')' - (\beta_{F_n}', \beta_{F_n,P}')')$ has the same limiting distribution under any sequence of local alternatives as it does when $F_n = F$ for all n. As argued by Ichimura and Newey (2017, Section 3, p.14), this condition is a mild one and it allows one to bypass imposing primitive conditions of asymptotic linearity and to focus on the main dominance result of this paper. Note that in (3.19), $\hat{\beta}_{n,P}$ is re-centered using δ_{F_n} and the presumption that $n^{1/2}\delta_{F_n} \to d$. Moreover, $V_{F,SP} \geq V_{F,P}$

states the intuition that imposing parametric restrictions generally leads to (weak) efficiency gain.[14] Formal justification of Condition 2(i) is in Appendix E, and the following is only a brief explanation of how this intuitive condition can be justified by Le Cam's third lemma (e.g., Van der Vaart, 2000, Example 6.7) and the definition of semiparametric efficiency bound (see Bickel, Klaassen, Ritov, and Wellner, 1993, Chapter 3). First, when $\|d\| = 0$, the parametric restrictions are correctly specified, due to Condition 1(i). So, the restricted nuisance function space $\mathcal{H}_g$ is a subspace of $\mathcal{H}$ that contains the true nuisance function h_F. Using an argument similar to that in the proof of Lemma 1 in Ackerberg et al. (2014), one can show that the semiparametric efficiency bound of the restricted model (with nuisance function space $\mathcal{H}_g$) is smaller than that of the unrestricted model (with nuisance function space $\mathcal{H}$),[15] because the latter is the supremum of all parametric submodels that include the former. In consequence, it is natural to require that $V_{F,SP} \geq V_{F,P}$ in this case.[16] Second, when $\|d\| < \infty$ but $\|d\| \neq 0$, the asymptotic variance-covariance matrix of $\hat{\beta}_{n,P}$ remains $V_{F,P}$ by the local regularity and Le Cam's third lemma. In addition, the asymptotic variance-covariance matrix of $\hat{\beta}_{n,SP}$ remains $V_{F,SP}$ regardless of the parametric restrictions. Therefore, $V_{F,SP} \geq V_{F,P}$ still holds. Condition 2(ii) is also intuitive since it states that when the parametric restrictions are severely misspecified, $\hat{\beta}_{n,P}$ will have an infinitely large asymptotic bias. Formal justification of Condition 2(ii) is also in Appendix E.

Condition 2 is a high-level condition that might be ensured by different primitive conditions in specific semiparametric models, on which there have been many important contributions (e.g., Robinson, 1988; Klein and Spady, 1993; Hirano et al., 2003; Cheng et al., 2019). Condition 2 bypasses those conditions and focuses on the common asymptotic properties in preparation for the discussion of the averaging estimator. Also note that Condition 2 takes the consistency of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ for respective (pseudo-)true values defined in (3.1) and (3.8) as presumption, for which the primitive conditions have been studied extensively (e.g., Newey and McFadden, 1994, Section 2).

Define

$$A_F \equiv \Upsilon(V_{F,SP} - C_F) \quad \text{and} \quad B_F \equiv \Upsilon(V_{F,SP} + V_{F,P} - 2C_F). \tag{3.20}$$

Given the high-level Condition 2, the following lemma follows immediately.

Lemma 1. *Suppose Conditions 1 and 2 hold. Also suppose that $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$ have finite probability limits.*

(i) If $\|d\| < \infty$, then

$$\hat{w}_n \xrightarrow{d.} w_F \equiv \frac{tr(A_F)}{tr(B_F) + (\xi_{F,P} + d - \xi_{F,SP})'\Upsilon(\xi_{F,P} + d - \xi_{F,SP})}, \tag{3.21}$$

[14] Condition 2(i) is easy to verify for a specific model. The direct verification of Condition 2(i) for the partially linear model of Section 4 is in Appendix D.

[15] That is, the difference between the two is a negative semi-definite matrix.

[16] Following the Ackerberg et al. (2014) approach, one needs to define another nuisance parameter η, which captures the features of the distribution of data Z other than those determined by β and h, then characterize the tangent space (see Newey, 1990; Bickel et al., 1993) for both the unrestricted and the restricted models. The efficient score function of β in each model is therefore the projection residual of the score function of β onto the model's tangent space. Since the unrestricted models include the restricted models as a subspace, the tangent space of the former includes that of the latter as a subspace as well. This implies that the efficient score function of β in the former has smaller norm than that in the latter. This in turn implies that the semiparametric efficiency bound of the former, which is the inverse of the squared norm of the efficient score function, is larger than that of the latter. Strictly speaking, it is still possible that the two step parametric estimator is asymptotically less efficient than the semiparametric estimator despite the opposite relative magnitude of their efficiency bounds, but since Crepon, Kramarz, and Trognon (1997) and Newey and Powell (1999) show in different models that the two step estimators achieve the efficiency bounds if the first step is exactly identified, the high-level condition $V_{F,SP} \geq V_{F,P}$ in Condition 2(i) does not go without justification.

which in turn implies that

$$n^{1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_F) \xrightarrow{d.} \bar{\xi}_{F,d} \equiv (1 - w_F)\xi_{F,SP} + w_F(\xi_{F,P} + d). \tag{3.22}$$

(ii) If $\|d\| = \infty$*, then* $\hat{w}_n \xrightarrow{p.} 0$ *and* $n^{1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_F) \xrightarrow{d.} \xi_{F,SP}$*.*

Proof. See Appendix A. ■

Remark 2. *Condition 2(i) assumes* $V_{F,SP} \geq V_{F,P}$ *because it is the case in which the averaging is meaningful (otherwise* $\hat{\beta}_{n,SP}$ *dominates* $\hat{\beta}_{n,P}$*, and hence no averaging is needed). Incorporating the possibility of* $V_{F,SP} < V_{F,P}$ *can be done by modifying the data-driven averaging weight in* (2.2) *to*

$$\tilde{w}_n \equiv \frac{tr[\Upsilon(\widehat{V}_{n,SP} - \widehat{C}_n)] \cdot \mathbb{I}\{\widehat{V}_{n,SP} \geq \widehat{V}_{n,P}\}}{tr[\Upsilon(\widehat{V}_{n,SP} + \widehat{V}_{n,P} - 2\widehat{C}_n)] + n(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})'\Upsilon(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})}. \tag{3.23}$$

When $V_{F,SP} \geq V_{F,P}$*, the weights* $\tilde{w}_n = \hat{w}_n$ *with probability one, so the asymptotic results in Lemma 1 still hold. Since the uniform dominance theory (uniform over* $\|d\|$ *values but not* $V_{F,SP}$ *or* $V_{F,P}$*) in Theorem 1 builds on these asymptotic results, it will not be affected by such modification of the weight. When* $V_{F,SP} < V_{F,P}$*, on the other hand, it is easy to see that the weight* $\tilde{w}_n$ *converges to* 0 *in probability. In this case, the results in Lemma 1(ii) (i.e.,* $n^{1/2}(\hat{\beta}_{n,\tilde{w}_n} - \beta_F) \xrightarrow{d.} \xi_{F,SP}$*) hold regardless of* $\|d\|$ *value, and the resulting averaging estimator* $\hat{\beta}_{n,\tilde{w}_n}$ *has the same asymptotic properties as* $\hat{\beta}_{n,SP}$*, only weakly dominating the latter (and hence the uniform dominance result still holds).*

The practical implication of this remark is that if a researcher is uncertain whether the condition $V_{SP} \geq V_P$ *holds, then the weight* (3.23) *and the resulting averaging estimator can be used.*

Condition 3. *Suppose* $\mathcal{F}$ *is such that the following holds.*

(i) $\mathcal{S}$ *is compact, with* $\mathcal{S}$ *defined in* (3.18)*;*

(ii) for any $F \in \mathcal{F}$ *such that* $\delta_F = 0$ *with* δ_F *defined in* (3.9)*, there exists a constant* $\epsilon_F > 0$ *such that for any* $\tilde{\delta} \in \mathbb{R}^k$ *with* $0 \leq \|\tilde{\delta}\| < \epsilon_F$*, there is* $\tilde{F} \in \mathcal{F}$ *with* $\delta_{\tilde{F}} = \tilde{\delta}$ *and* $\|\bar{S}(\tilde{F}) - \bar{S}(F)\| \leq C\|\tilde{\delta}\|^\kappa$ *for some* $C, \kappa > 0$*, where* $\bar{S}(F)$ *is defined in* (3.17)*.*

Condition 3(i) is necessary for applying the subsequence argument to show the uniform dominance result. Recall that $\mathcal{S}$ defined in (3.18) is a subset of a finite-dimensional Euclidean space, so Condition 3(i) is equivalent to $\mathcal{S}$ being bounded and closed. vech($V_{F,SP}$), vech($V_{F,P}$) and vec(C_F) are bounded if both $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are locally regular estimators, which is implied by Condition 2(i) for $\|d\| < \infty$ (see the discussion after Condition 2 for details). $\mathcal{S}$ being closed is not restrictive in the sense that if $\mathcal{S}$ is not closed, then one can define it to be the closure of $\mathcal{S}$ and the main uniform dominance result still holds. Condition 3(ii) says that for any $F \in \mathcal{F}$ satisfying the parametric restrictions, there are many DGPs $\tilde{F} \in \mathcal{F}$ that are close to F, where the closeness of two DGPs is measured by the distance between $\bar{S}(\tilde{F})$ and $\bar{S}(F)$.[17] This condition will be used in the subsequence argument to show the uniform dominance and is not restrictive, since it means that the DGP set $\mathcal{F}$ is rich enough, which makes the uniform dominance result harder to hold.

Once a specific model is given, Conditions 1 and 3 can be verified directly, and the literature often has developed primitive conditions for Condition 2. Appendix D details the primitive conditions of Condition 2 for the partially linear model in Section 4 and verifies Conditions 1 - 3 for the parameterization used in the Monte Carlo experiments in that section.

[17] Under Condition 3(ii), for any $F \in \mathcal{F}$ with $\delta_F = 0$ and any sequence of DGPs $\{F_n\}_{n=1}^\infty$ such that $n^{1/2}\delta_{F_n} \to d$ with $\|d\| < \infty$, there exists a sequence of DGPs $\{\tilde{F}_n\}_{n=1}^\infty$ satisfying the requirement of Condition 2(i), and hence the convergence result in (3.19) holds. This interpretation is related to Assumptions A and B in Andrews and Guggenberger (2010) and Assumptions A0 and B0 in Andrews and Guggenberger (2009). The author thanks the co-editor for pointing this out.

In order to state an important intermediate result and to explain its rationale, some additional notation is needed. For any $F \in \mathcal{F}$ and any $d \in \mathbb{R}^k_\infty$, define

$$u_{F,d} \equiv (d', \text{vech}(V_{F,SP})', \text{vech}(V_{F,P})', \text{vec}(C_F)')'. \tag{3.24}$$

Note that the subvector d of $u_{F,d}$ does not depend on F, and the rest of $u_{F,d}$ does not depend on d. Let

$$\mathcal{U} \equiv \{u_{F,d}\text{: } \|d\| < \infty, \text{ and } F \in \mathcal{F} \text{ with } \delta_F = 0\}, \tag{3.25}$$

and

$$\mathcal{U}_\infty \equiv \{u_{F,d}\text{: } \|d\| = \infty, \text{ and } F \in \mathcal{F}\}. \tag{3.26}$$

For any $u_{F,d} \in \mathcal{U} \cup \mathcal{U}_\infty$, define

$$r(u_{F,d}) \equiv \begin{cases} \mathbb{E}_F\left(\bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} - \xi'_{F,SP} \Upsilon \xi_{F,SP}\right), & \text{if } u_{F,d} \in \mathcal{U}; \\ 0, & \text{if } u_{F,d} \in \mathcal{U}_\infty, \end{cases} \tag{3.27}$$

where $\bar{\xi}_{F,d}$ and $\xi_{F,SP}$ are defined in (3.22) and (3.19), respectively. $\mathcal{U}$ and $\mathcal{U}_\infty$ defined here may appear similar to the set $\mathcal{S}$ defined in (3.18), but they are different. For any $u_{F,d} \in \mathcal{U} \cup \mathcal{U}_\infty$, the corresponding $\delta \equiv n^{-1/2}d$ is a different object from δ_F associated with F. $\mathcal{S}$ is the set of *actual* nuisance parameter vectors that determine the asymptotic properties of $\hat{\beta}_{n,SP}$, $\hat{\beta}_{n,P}$ and $\hat{\beta}_{n,\hat{w}_n}$ under DGPs in $\mathcal{F}$. In contrast, $\mathcal{U}$ is the set of all *hypothetical* nuisance parameter vectors that would have prevailed had the asymptotic variance-covariance matrices $V_{F,SP}$, $V_{F,P}$ and C_F been the same as some DGP with zero bias ($\delta_F = 0$) from $\mathcal{F}$ and had the asymptotic bias d been finite. Note that if $u_{F,d} \in \mathcal{U}$ (i.e., $\|d\| < \infty$), the corresponding δ ranges from being zero to approaching to infinity at any rate that is not faster than $n^{1/2}$, corresponding to correct specification or mild misspecification of the parametric restrictions. Similarly, $\mathcal{U}_\infty$ is the set of all *hypothetical* nuisance parameter vectors that would have prevailed had the asymptotic variance-covariance matrices $V_{F,SP}$, $V_{F,P}$ and C_F been the same as some DGP from $\mathcal{F}$ and had the asymptotic bias d been infinite. Note that if $u_{F,d} \in \mathcal{U}_\infty$ (i.e., $\|d\| = \infty$), the corresponding δ approaches to infinity at faster than $n^{1/2}$ rate, corresponding to severe misspecification of the parametric restrictions. Together, $\mathcal{U}$ and $\mathcal{U}_\infty$ are a device that allows one to compare the asymptotic risk of $\hat{\beta}_{n,\hat{w}_n}$ to that of $\hat{\beta}_{n,SP}$ uniformly over varied degrees of misspecification of the parametric restrictions.

To show the main uniform dominance result, the following lemma implies that one can first approximate the bounds of asymptotic risk difference using $r(u_{F,d})$ for $u_{F,d} \in \mathcal{U}$ and for $u_{F,d} \in \mathcal{U}_\infty$ separately, and then combine the two cases together.

Lemma 2. *Suppose: (i) Conditions 1 - 3 hold; (ii) $tr(A_F) > 0$ and $tr(B_F) > 0$, where A_F and B_F are defined in* (3.20).[18] *Then*

$$Asy\overline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \max\left\{\sup_{u_{F,d} \in \mathcal{U}} r(u_{F,d}), 0\right\} \tag{3.28}$$

$$Asy\underline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \min\left\{\inf_{u_{F,d} \in \mathcal{U}} r(u_{F,d}), 0\right\}. \tag{3.29}$$

Proof. See Appendix A. ■

[18] As shown in Appendix A (Lemmas A.3 and A.4), a weaker condition than (ii) – $\text{tr}(A_F) \geq 0$ and $\text{tr}(B_F) > 0$ – is sufficient for proving Lemma 2. Due to the definitions of A_F and B_F, however, if $V_{F,SP} \geq V_{F,P}$ as postulated in Condition 2(i), then $\text{tr}(B_F) > 0$ implies $\text{tr}(A_F) > 0$.

If the parametric restrictions are severely misspecified, then one has $u_{F,d} \in \mathcal{U}_\infty$ (and hence $\|d\| = \infty$). In this case, Lemma 1(ii) states that the asymptotic distributions of $\hat{\beta}_{n,\hat{w}_n}$ and $\hat{\beta}_{n,SP}$ are the same, and therefore $r(u_{F,d}) = 0$. The key message of Lemma 2 is that the upper (or lower) bound of the asymptotic risk difference is determined by the maximum between $\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d})$ and $\sup_{u_{F,d}\in\mathcal{U}_\infty} r(u_{F,d}) = 0$ (or the minimum between $\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d})$ and $\inf_{u_{F,d}\in\mathcal{U}_\infty} r(u_{F,d}) = 0$).

By Lemma 2, showing that $\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}) \leq 0$ and $\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d}) < 0$ is sufficient for the following uniform dominance theorem.

Theorem 1. *Suppose Conditions 1 - 3 hold. Let A_F and B_F be those matrices defined in* (3.20)*, and let $\rho_{\max}(\cdot)$ denote the largest eigenvalue of a square matrix. If $tr(A_F) > 0$, $tr(B_F) > 0$ and $tr(A_F) \geq 4\rho_{\max}(A_F)$ for any $F \in \mathcal{F}$ with $\delta_F = 0$, then* (3.15) *and* (3.16) *hold; that is, the averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ uniformly dominates the semiparametric estimator $\hat{\beta}_{n,SP}$.*

Proof. See Appendix A. ■

To give some intuition for the conditions of the dominance result in Theorem 1, consider the case where the researcher chooses $\Upsilon = (V_{F,SP} - C_F)^{-1}$. In this case, the presumption $V_{F,SP} \geq V_{F,P}$ and the invertibility of $V_{F,SP} - C_F$ requires that $V_{F,SP} > C_F$, which is a necessary condition for $V_{F,SP} > V_{F,P}$. The latter indicates that the parametric estimator should achieve *strict* efficiency gain over the semiparametric estimator. In addition, the condition $\text{tr}(A_F) \geq 4\rho_{\max}(A_F)$ becomes $k \geq 4$, which requires the researcher to consider the overall risk of multiple parameters of interest, but not a single coordinate. Such a dimension condition is common for shrinkage estimators. For example, the condition here is stronger than the condition $k \geq 3$ for the estimators in James and Stein (1961) and Hansen (2016), the same as $k \geq 4$ for the estimator in Cheng et al. (2019), and is weaker than $k \geq 5$ for the estimators in Judge and Mittelhammer (2004) and Mittelhammer and Judge (2005).

The averaging weight $\hat{w}_n$ in (2.2) is a sample analog of the infeasible optimal weight under the quadratic loss function in (2.1). To see this, note that by Condition 2(i), for any fixed weight w, the asymptotic distribution of $\hat{\beta}_{n,w}$ when $\|d\| < \infty$ is obtained by the continuous mapping theorem:

$$n^{1/2}(\hat{\beta}_{n,w} - \beta_F) \xrightarrow{d.} \xi_{F,w} \equiv (1-w)\xi_{F,SP} + w(\xi_{F,P} + d).$$

Since the asymptotic risk, defined in (2.1), is quadratic in w, the optimal weight w^* that minimizes the asymptotic risk under DGP F is

$$w^* \equiv \frac{\text{tr}[\Upsilon(V_{F,SP} - C_F)]}{\text{tr}[\Upsilon(V_{F,SP} + V_{F,P} - 2C_F)] + d'\Upsilon d}.$$

If $\hat{\beta}_{n,P}$ is asymptotically efficient under the parametric restrictions, then $C_F = V_{F,P}$ and the optimal weight simplifies to

$$w^* = \frac{\text{tr}[\Upsilon(V_{F,SP} - V_{F,P})]}{\text{tr}[\Upsilon(V_{F,SP} - V_{F,P})] + d'\Upsilon d}.$$

Further note that $n^{1/2}(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})$ is an asymptotically unbiased estimator of d when $\|d\| < \infty$, so the averaging weight $\hat{w}_n$ in (2.2) is a sample analog of w^*.

When $\|d\| = \infty$, the parametric estimator $\hat{\beta}_{n,P}$ is so severely biased that a sensible averaging estimator ought to assign it zero weight. This intuition is echoed by Condition 2(ii) and Lemma 1(ii), which imply that the feasible averaging weight given in (2.2) approaches to zero.

It is worth pointing out that because $n^{1/2}(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})$ is only asymptotically unbiased for d but not consistent,[19] and w_F in (3.21) is a random variable and in general not unbiased for w^* in light of the Jensen's inequality, so $\hat{w}_n$ is neither a consistent nor an unbiased estimator for the infeasible optimal weight w^*. Proving the uniform dominance of the averaging estimator, therefore, is more challenging than it might appear at first sight, since $\hat{\beta}_{n,SP}$, $\hat{\beta}_{n,P}$ and $\hat{w}_n$ are mutually dependent random variables and their randomness needs to be dealt with at the same time. For this reason, the subsequence technique in Cheng et al. (2019) is necessary for proving Theorem 1.

Inference. The inference of averaging estimators generally differs from standard estimators, in that the averaging weights often are random variables that correlate with the candidate estimators, which renders the asymptotic distribution of $\hat{\beta}_{n,\hat{w}_n}$ non-standard. Fortunately, inference can still be made here by adapting a conservative two-step inference method proposed by Claeskens and Hjort (2008, Section 7.5.4).

To adapt Claeskens and Hjort (2008)'s two-step inference method to the averaging estimator in this paper, first note that given any fixed finite d, Condition 2(i) and Lemma 1(i) tell us that $\xi_{F,SP}$ and $\xi_{F,P}$ completely determine the joint asymptotic distributions of $\hat{\beta}_{n,SP}$, $\hat{\beta}_{n,P}$ and $\hat{w}_n$. So, they also determine the asymptotic distribution of $\hat{\beta}_{n,\hat{w}_n}$, represented by $\bar{\xi}_{F,d}$ in (3.22). As a result, for fixed finite d and any confidence level $1-\alpha_2$, the confidence set of $\Delta_n \equiv \sqrt{n}(\hat{\beta}_{n,\hat{w}_n} - \beta_F)$, denoted as $CI_{1-\alpha_2}(\Delta_n|d, \widehat{V})$, can be constructed by simulating (S number of random draws for a given large number S) from the joint distribution of $\xi_{F,SP}$ and $\xi_{F,P}$ provided in Condition 2(i) and picking the upper and lower $\alpha_2/2$ critical values of the simulated $\bar{\xi}_{F,d}$ values given in (3.22).[20]

To account for the fact that d is unknown, note that Condition 2(i) immediately implies that $n^{1/2}(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP}) \xrightarrow{d.} \mathcal{N}(d, V_{F,P} + V_{F,SP} - C_F - C_F')$. This enables one to construct the following confidence set of d for any confidence level $1-\alpha_1$:

$$CI_{1-\alpha_1}(d|\hat{\beta}, \widehat{V}) \equiv \{d:\ n(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP} - d)'\widehat{V}_d^{-1}(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP} - d) \leq \chi^2_{1-\alpha_1}(k)\}, \tag{3.30}$$

where $\widehat{V}_d \equiv \widehat{V}_{n,P} + \widehat{V}_{n,SP} - \widehat{C}_n - \widehat{C}_n'$ and $\chi^2_{1-\alpha_1}(k)$ is the $1-\alpha_1$ quantile of the χ^2 distribution with degrees of freedom k.

In summary, the two-step method proceeds as follows:

(1) For any confidence level $1-\alpha$, pick α_1 and α_2 such that $\alpha_1 + \alpha_2 = \alpha$, and construct the $1-\alpha_1$ confidence set $CI_{1-\alpha_1}(d|\hat{\beta}, \widehat{V})$ of d, defined in (3.30).

(2) For each $d \in CI_{1-\alpha_1}(d|\hat{\beta}, \widehat{V})$, construct the $1-\alpha_2$ confidence set $CI_{1-\alpha_2}(\Delta_n|d, \widehat{V})$ of Δ_n via simulation described above, then take the union $\cup_{d \in CI_{1-\alpha_1}(d|\hat{\beta}, \widehat{V})} CI_{1-\alpha_2}(\Delta_n|d, \widehat{V})$.

That is, for chosen α_1 and α_2 such that $\alpha_1 + \alpha_2 = \alpha$, the $1-\alpha$ confidence set of β_F is just

$$CI_{1-\alpha}(\beta|\hat{\beta}, \widehat{V}) \equiv \{\beta:\ \sqrt{n}(\hat{\beta}_{n,\hat{w}_n} - \beta) \in CI_{1-\alpha_2}(\Delta_n|d, \widehat{V}),\ \text{for some } d \in CI_{1-\alpha_1}(d|\hat{\beta}, \widehat{V})\}. \tag{3.31}$$

In practice, this union can be well approximated by taking a large number of d values satisfying (3.30), and by taking the union of the resulting sets $CI_{1-\alpha_2}(\sqrt{n}(\hat{\beta}_{n,\hat{w}_n} - \beta_F)|d, \widehat{V})$ over all such d values. Such a union set allows one to make inference about β_F based on the data.

The next lemma follows Claeskens and Hjort (2008, Section 7.5.4) and Kitagawa and Muris (2016, Appendix A) to show that the confidence set $CI_{1-\alpha}(\beta_F|\hat{\beta}, \widehat{V})$ is asymptotically valid. Note that the validity

[19] In fact, d is not root-n estimable, since its information bound is zero.

[20] In simulating from the joint distribution in Condition 2(i), one obviously needs to replace the unknown variance-covariance matrices with their consistent estimators. Here and in the rest of this subsection, $\hat{\beta}$ is a shorthand for $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$, and $\widehat{V}$ is a shorthand for $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$.

of (3.32) below does not depend on the value d, so the confidence interval $CI_{1-\alpha}(\beta|\hat{\beta}, \widehat{V})$ is uniformly valid regardless of the degree of misspecification.

Lemma 3. *Suppose Conditions 1 - 3 hold. Let α_1 and α_2 be chosen non-negative numbers such that $\alpha_1 + \alpha_2 = \alpha$, then*

$$\lim_{n\to\infty} P_F\left(\beta_F \in CI_{1-\alpha}(\beta|\hat{\beta}, \widehat{V})\right) \geq 1 - \alpha. \tag{3.32}$$

Proof. See Appendix A. ■

In contrast to this two-step method, a naive inference method based on the averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ might treat the averaging weight $\hat{w}_n$ as non-random and compute the asymptotic variance-covariance matrix of the averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ as $\hat{w}_n^2\widehat{V}_{n,P} + (1-\hat{w}_n)^2\widehat{V}_{n,SP} + \hat{w}_n(1-\hat{w}_n)\widehat{C}_n + \hat{w}_n(1-\hat{w}_n)\widehat{C}_n'$. In addition, standard inference based only on the semiparametric estimator $\hat{\beta}_{n,SP}$ is always feasible. Section 4 will compare the finite sample sizes and powers of the two-step method with the naive method (two variations) and the standard $\hat{\beta}_{n,SP}$-based inference method in a partially linear model example.

4 An Example with Monte Carlo Experiments

Example - partially linear model. *One is interested in estimating β in a partially linear model*

$$Y_i = X_{1i}'\beta + s(X_{1i}, X_{2i}) + U_i, \tag{4.1}$$

where $\mathbb{E}(U_i|X_{1i}, X_{2i}) = 0$, X_{1i} is a $k \times 1$ vector, X_{2i} is an $l \times 1$ vector, and X_{1i} and X_{2i} are assumed not to overlap for simplicity. The identification of β requires that $s(X_{1i}, X_{2i})$ and X_{1i} are not perfectly collinear.[21]

The estimator of β that results from $s(x_1, x_2)$ being approximated by a series of basis functions (e.g., polynomials) that increases with the sample size is one example of the semiparametric estimator $\hat{\beta}_{n,SP}$.[22] *If one imposes certain parametric-form restriction on $s(x_1, x_2)$ – for example, $s(x_1, x_2)$ is a linear function of x_2 only – then the usual least squares estimator of β could serve as one example of the parametric estimator $\hat{\beta}_{n,P}$.*[23]

Although the semiparametric models considered in this paper are flexible enough to include many examples, this partially linear model is put in the spotlight because it highlights a few distinct features of the averaging estimator in this paper. First, unlike in Cheng et al. (2019), the parametric estimator $\hat{\beta}_{n,P}$ in this paper (e.g., the least squares estimator in this example) need not be asymptotically efficient under the parametric restrictions. Second, the asymptotic distribution of a two step M estimator generally depends on the *presence* of the first step nuisance parameter and how it is *modeled* (e.g., parametrically or nonparametrically), even in the absence of first step *estimation error* like the partially linear model.[24] Third, the Stein-type condition amounts to a dimensionality condition ($k \geq 4$) for $\Upsilon = (V_{F,SP} - V_{F,P})^{-1}$ (discussed after Theorem 1) and it can be easily fulfilled in this example.

[21] To be precise, $\mathbb{E}\{[X_{1i} - \mathbb{E}[X_{1i}|s(X_{1i}, X_{2i})] \cdot [X_{1i} - \mathbb{E}[X_{1i}|s(X_{1i}, X_{2i})]'\}$ is positive definite.

[22] Many semiparametric estimators of β in partially linear models have been proposed in the literature (e.g., Robinson, 1988; Donald and Newey, 1994). In particular, since partially linear models may arise as a "reduced form" of the sample selection models (see discussion on pages 5-8 of Ahn and Powell, 1993, and reference therein), many semiparametric estimators of β in sample selection models (with potentially nonparametric selection equation) have been proposed and examined under various identification conditions, for example, Gallant and Nychka (1987); Newey, Powell, and Walker (1990); Ahn and Powell (1993); Newey (2009). They could all serve as the $\hat{\beta}_{n,SP}$ in this paper, provided that the conditions in Section 3 are satisfied.

[23] Least squares estimator is generally considered as a semiparametric estimator, since the distribution of U_i is usually left unspecified. If the distribution of U_i is parametrically specified, then the resulting least squares estimator is truly a parametric estimator. In this example, however, both are referred to as "parametric estimators" because they both impose certain restrictions on the nuisance function $s(x_1, x_2)$.

[24] Appendix D shows that for this partially linear model, the influence functions of $\hat{\beta}_{n,P}$ and $\hat{\beta}_{n,SP}$ ((D.2) and (D.3), respectively) differ, although neither of them contains the correction term of the first step estimation error.

Monte Carlo experiments. In the Monte Carlo experiments, the interest is to estimate $\beta \equiv (\beta_1, \beta_2, \beta_3, \beta_4)'$ in the following parameterization of the model in (4.1)

$$Y = \sum_{j=1}^{4} \beta_j X_{1j} + \sum_{j=1}^{4} \theta_{1j} X_{2j} + \rho \left(\sum_{j=1}^{4} \theta_{2j} \exp(X_{2j}) + \sum_{j=1}^{4} \theta_{3j} X_{1j} X_{2j} \right) + U, \tag{4.2}$$

where X_{1j} and X_{2j} denote the jth coordinate of X_1 and X_2, respectively ($k = 4$ here).

The parametric estimator $\hat{\beta}_{n,P}$ results from the following (misspecified) linear regression

$$Y = \theta_0 + \sum_{j=1}^{4} \beta_j X_{1j} + \sum_{j=1}^{4} \theta_{1j} X_{2j} + V, \tag{4.3}$$

while the semiparametric series estimator $\hat{\beta}_{n,SP}$ results from a linear regression of Y on X_1 and polynomials of X_1 and X_2 which exclude linear functions of X_1 (as proposed by Donald and Newey, 1994).[25] When $\rho \neq 0$, the parametric estimator $\hat{\beta}_{n,P}$ suffers from the familiar "omitted variable bias", since the term in the bracket in (4.2) generally correlates with both X_1 and X_2.

In the experiments, U is independent of $(X_1', X_2')'$ and randomly drawn from $\mathcal{N}(0, 0.5^2)$, and $(X_1', X_2')'$ is randomly drawn from $\mathcal{N}(2 \times \ell_8, V_X)$ with

$$V_X = \begin{bmatrix} 0.5^2 \times I_4 & 0.05 \times L_{4\times 4} \\ 0.05 \times L_{4\times 4} & 0.5^2 \times I_4 \end{bmatrix}, \tag{4.4}$$

where ℓ_8 is an 8×1 vector of ones, I_4 is the 4×4 identity matrix and $L_{4\times 4}$ is a 4×4 matrix of ones. The parameter values are $\beta = (4, 3, 2, 1)'$, $\theta_1 = (1, 1, 1, 1)'$, $\theta_2 = (1, 2, 3, 4)'$ and $\theta_3 = (5, 6, 7, 8)'$. The value of ρ, which determines the degree of misspecification of $\hat{\beta}_{n,P}$, varies from 0 to 1.3 with 0.05 step width. The sample size is $n = 1000$, and the number of Monte Carlo replications is $R = 10000$.[26] The weighting matrix is $\Upsilon = I_4$, so that the risk function is the MSE.

The MSEs, squared biases and variances of $\hat{\beta}_{n,SP}$, $\hat{\beta}_{n,P}$ and $\hat{\beta}_{n,\hat{w}_n}$ are plotted against the degree of misspecification ρ in Figure 1. The MSEs are normalized by those of $\hat{\beta}_{n,SP}$, which is represented by the thin solid black line at unity. So, the MSE of an estimator being below this unity benchmark means that it has smaller MSEs than $\hat{\beta}_{n,SP}$. The normalized MSEs of $\hat{\beta}_{n,\hat{w}_n}$ with the averaging weight $\hat{w}_n$ computed using the influence-function-based asymptotic variance-covariance matrix estimates (detailed in Appendix C) are represented by the thick solid black line, while those using the bootstrapped asymptotic variance-covariance matrix estimates ($B = 200$ bootstrap replications) are represented by the dashed black line.[27] The normalized MSEs of $\hat{\beta}_{n,P}$ are represented by the dash-dotted black line. The squared biases and variances of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$, not normalized, are plotted as well to facilitate the understanding of the performance of the estimators.[28] The squared biases of $\hat{\beta}_{n,SP}$ are represented by the dashed yellow line with triangle markers, and those of $\hat{\beta}_{n,P}$ by the dashed blue line with round markers; the latter increases so quickly with ρ and shoots outside the figure range before ρ reaches 0.1. The variances of $\hat{\beta}_{n,SP}$ are represented by the dotted yellow line with triangle markers, and those of $\hat{\beta}_{n,P}$ by the dotted blue line with round markers; the former remains stable around the level of 45 and is outside the figure range.

Figure 2 plots the Monte Carlo distributions (kernel densities) of the first coordinate of the averaging

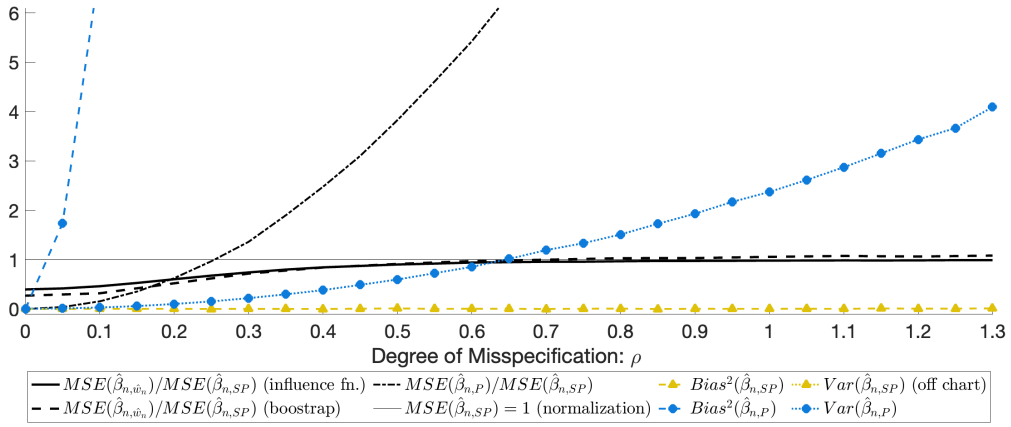
[25] Based on a leave-one-out cross validation procedure performed on a preliminary sample, polynomials up to the fourth order are used for $\hat{\beta}_{n,SP}$ in the Monte Carlo experiments.

[26] Alternative sample sizes $n = 100, 250, 500$ are also considered, and the results are similar (not reported).

[27] To save time, the bootstrap averaging estimator is only computed $R = 1000$ replications with 0.1 step width of ρ values.

[28] The author thanks an anonymous referee for suggesting plotting them and the distributions of the averaging weights.

Figure 1: Monte Carlo MSE, Bias^2 and Variance of $\hat{\beta}_{n,SP}$, $\hat{\beta}_{n,P}$ and $\hat{\beta}_{n,\hat{w}_n}$ for the Partially Linear Model



Notes: (1) Bootstrap results are based on $R = 1000$ Monte Carlo replications, $n = 1000$ sample size and $B = 200$ bootstrap replications.
(2) All other results are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size.
(3) MSEs are normalized by dividing those of the semiparametric estimator $\hat{\beta}_{n,SP}$.
(4) Squared biases and variances are not normalized by the MSEs of $\hat{\beta}_{n,SP}$. $Var(\hat{\beta}_{n,SP})$ is off the chart and around 45.
(5) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

estimator $\hat{\beta}_{n,\hat{w}_n}$ (thick solid lines) for representative ρ values. In the same figures, the normal distributions based on the naive inference method with the common standard error are represented by the thick dashed lines (one randomly chosen Monte Carlo replication) and dotted lines (averaged over all Monte Carlo replications). It is obvious that the naive inference method underestimates the randomness in the averaging estimator $\hat{\beta}_{n,\hat{w}_n,1}$, since it treats the averaging weight $\hat{w}_n$ as non-random.

Figure 3 plots the kernel densities of the averaging weight $\hat{w}_n$ for representative ρ values. The solid lines are based on the influence functions in Appendix C, and the dashed lines are based on the bootstrapping in Section 2. The difference between the two is undiscernible. As ρ value increases, both distributions of the averaging weight concentrate more and more towards one, confirming the results of Lemma 1.

Table 1 reports for different ρ values the rejection rates of $\hat{\beta}_{n,SP}$ with the common standard error and those of $\hat{\beta}_{n,\hat{w}_n}$ with both the naive and the two-step inference methods ($S = 1000$ random draws in the second step) for β_1, the first coordinate of β. Two variations of the naive inference method for $\hat{\beta}_{n,\hat{w}_n}$ are considered. The "Naive" one uses the common estimators of $V_{F,P}$ and C_F when computing the standard error, but they can be biased under misspecification (see the discussion after (C.3)). The "Naive (robust SE)" one uses the robust influence function (C.2) - (C.3) when computing the standard error (see Appendix (D) for details). For the "Size" columns, the test value is 4, the true value of β_1; for the "Power" columns, the test value is 0. Table 1 also reports the average ratios between the lengths of the two-step confidence intervals of $\hat{\beta}_{n,\hat{w}_n,1}$ and of the standard confidence intervals of $\hat{\beta}_{n,SP,1}$.

A few observations can be made about the Monte Carlo results. Firstly, regardless of the degree of misspecification, $\hat{\beta}_{n,SP}$ has almost zero bias but very large and stable variance, while $\hat{\beta}_{n,P}$ has much smaller variance but rapidly increasing bias. Secondly and consequently, the normalized MSEs of $\hat{\beta}_{n,P}$, compared to those of $\hat{\beta}_{n,SP}$, start from a negligible level and blow up quickly off the chart as ρ increases beyond 0.6. Thirdly, on the contrary, the normalized MSEs of $\hat{\beta}_{n,\hat{w}_n}$ stay below the unity benchmark regardless of the degree of misspecification, confirming the uniform asymptotic dominance result in Theorem 1. Fourthly,

both the influence function and the bootstrapping approaches lead to almost identical distributions of the averaging weights in Figure 3. The normalized MSEs of the two averaging estimators in Figure 1 differ, but to a very small degree. Fifthly, the asymptotic distributions based on the naive inference method with the common standard error in Figure 2 badly approximate the actual Monte Carlo distributions of the averaging estimator $\hat{\beta}_{n,\hat{w}_n,1}$ for all ρ values. Sixthly, both naive inference methods, with or without the robust standard error, lead to almost identical sizes and powers in Table 1, exhibiting significant size distortion (over-rejection). The two-step inference method, on the other hand, controls the size well and possesses decent powers. Finally, although the confidence interval based on $\hat{\beta}_{n,\hat{w}_n}$ with the two-step method is much longer than that based on $\hat{\beta}_{n,SP}$ with the common standard error, the two display comparable powers.

Figure 2 and Table 1 only present the Monte Carlo results for β_1, the first coordinate of β. Similar results for the other three coordinates are reported in Figures D.1 - D.3 and Tables D.1 - D.3 in Appendix D.

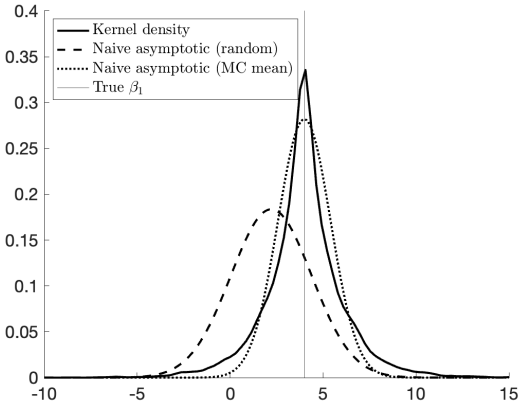
5 Conclusion

In a two step M estimation framework, this paper proposes a new estimator that averages between a semiparametric robust estimator and another one obtained under restricted first step, allowing the specific information used in developing the latter to complement the former. The subsequence technique employed in proving the uniform dominance of the averaging estimator is novel in the literature. The generality of the theoretical framework and the easy-to-implement computation and inference methods permit wide use of the proposed averaging estimator in semiparametric models.

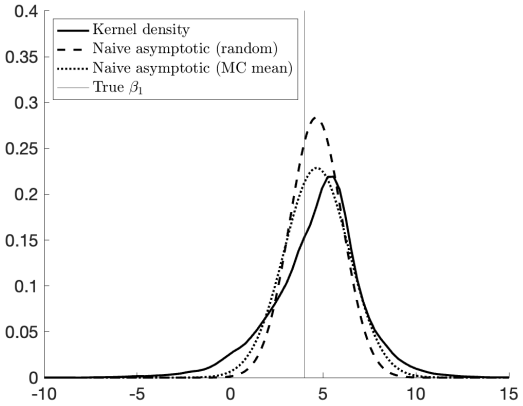
Inference based on the averaging estimator remains a challenging problem, as sharper uniformly valid confidence sets are desirable. The possibility of averaging the semiparametric estimator with more than one restricted estimator is also left for future research.

Figure 2: True vs. Naive Distributions of $\hat{\beta}_{n,\hat{w}_n,1}$ for the Partially Linear Model

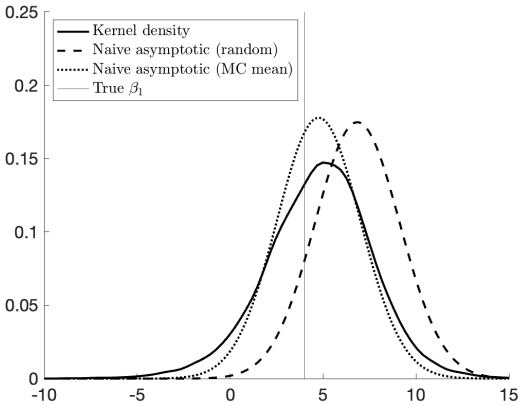
(a) Correct Specification ($\rho = 0$):



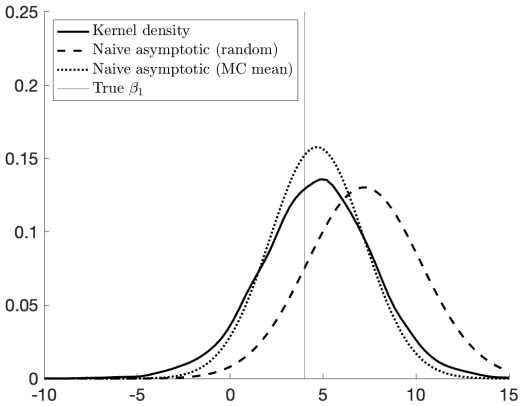
(b) Misspecification ($\rho = 0.2$):



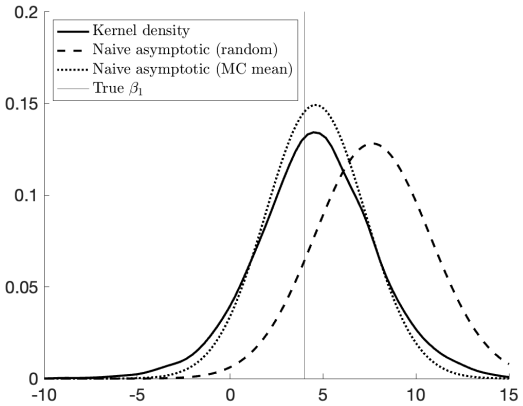
(c) Misspecification ($\rho = 0.4$):



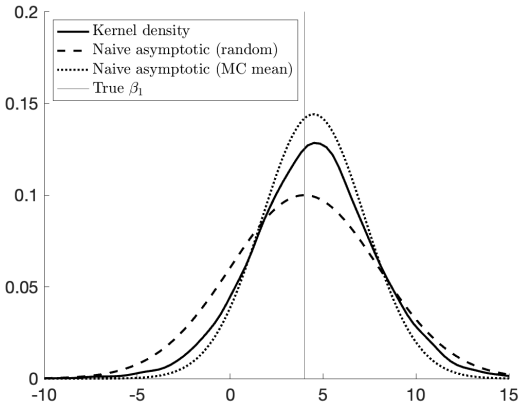
(d) Misspecification ($\rho = 0.6$):



(e) Misspecification ($\rho = 0.8$):



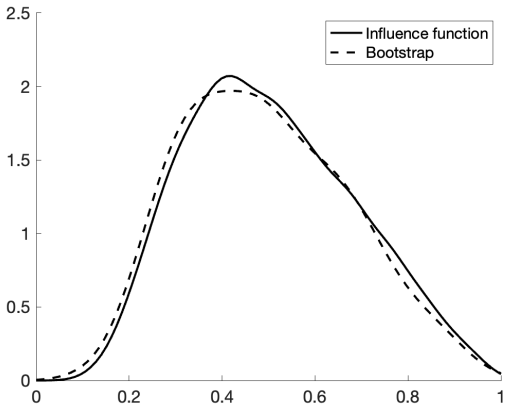
(f) Misspecification ($\rho = 1$):



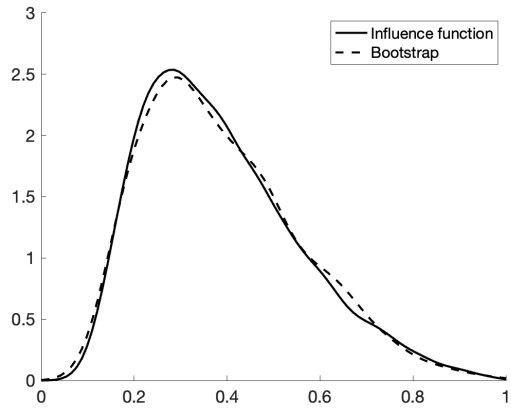
Notes: (1) All distributions are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size.
(2) Solid lines represent the MC distributions of $\hat{\beta}_{n,\hat{w}_n,1}$, the averaging estimator of β_1. Dashed and dotted lines both represent the asymptotic distribution of $\hat{\beta}_{n,\hat{w}_n,1}$ if the naive inference method, which takes $\hat{w}_n$ as fixed, is used. The former show a randomly chosen MC replication, while the latter show the average over all MC replications.
(3) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Figure 3: Distributions of $\hat{w}_n$ for the Partially Linear Model

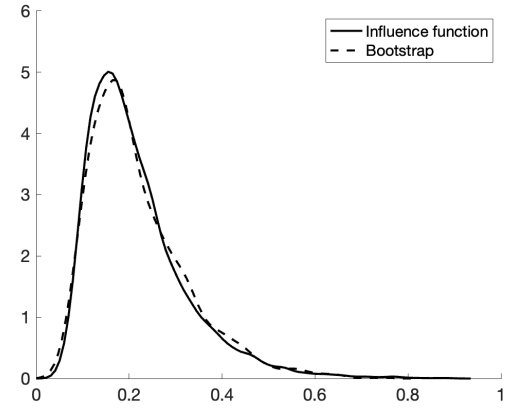
(a) Correct Specification ($\rho = 0$):



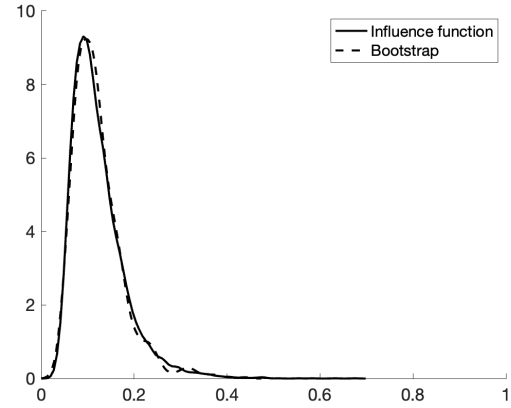
(b) Misspecification ($\rho = 0.2$):



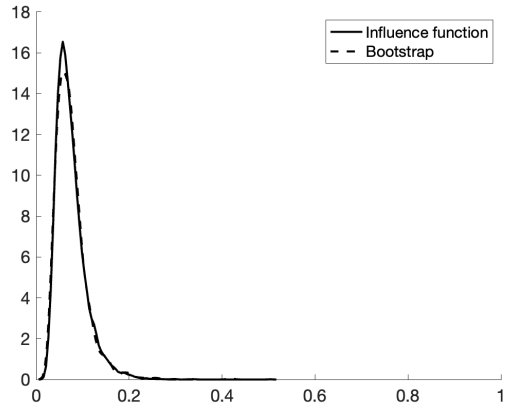
(c) Misspecification ($\rho = 0.4$):



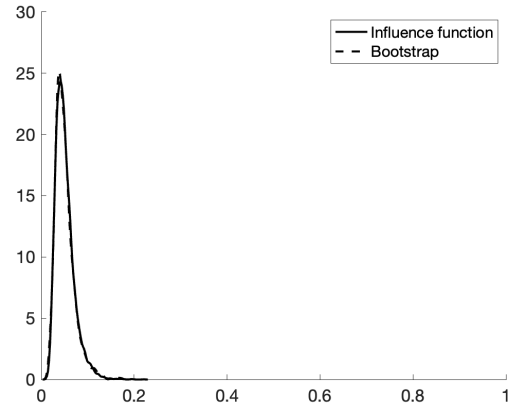
(d) Misspecification ($\rho = 0.6$):



(e) Misspecification ($\rho = 0.8$):



(f) Misspecification ($\rho = 1$):



Notes: (1) All distributions are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size.
(2) The distributions of the averaging weight $\hat{w}_n$ concentrate towards zero as ρ, the degree of misspecification, increases.
(3) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Table 1: Rejection Rates for $\hat{\beta}_{n,\hat{w}_n,1}$ in the Partially Linear Model (5% Level)

ρ	$\hat{\beta}_{n,SP,1}$		$\hat{\beta}_{n,\hat{w}_n,1}$						CI Length
			Naive		Naive (robust SE)		Two-step		$\frac{CI(\hat{\beta}_{n,\hat{w}_n,1})}{CI(\hat{\beta}_{n,SP,1})}$
	Size	Power	Size	Power	Size	Power	Size	Power	
0.00	9.27%	34.14%	9.30%	76.25%	9.16%	76.19%	1.65%	21.60%	32.8575
0.05	9.44%	34.64%	11.82%	76.76%	11.63%	76.70%	1.87%	24.64%	32.8520
0.10	9.62%	34.05%	16.53%	75.50%	16.32%	75.45%	2.01%	28.56%	32.8761
0.15	9.61%	35.34%	19.23%	73.95%	19.10%	73.93%	2.00%	32.91%	32.9347
0.20	9.53%	35.15%	19.98%	71.49%	19.83%	71.42%	2.48%	35.41%	33.0424
0.25	9.70%	35.66%	18.85%	68.32%	18.79%	68.26%	3.00%	37.26%	33.1656
0.30	9.97%	34.94%	17.77%	63.83%	17.71%	63.80%	3.32%	37.64%	33.3325
0.35	9.25%	34.24%	16.04%	61.12%	16.01%	61.10%	3.64%	36.88%	33.4862
0.40	9.93%	34.98%	15.88%	59.04%	15.85%	59.05%	4.56%	37.23%	33.6567
0.45	9.64%	34.72%	14.43%	56.00%	14.42%	56.00%	4.80%	37.08%	33.8317
0.50	9.67%	35.08%	13.42%	53.69%	13.42%	53.70%	5.13%	36.88%	33.9892
0.55	9.51%	34.70%	12.85%	51.87%	12.81%	51.81%	5.19%	36.13%	34.1402
0.60	9.37%	34.89%	11.86%	50.27%	11.82%	50.25%	5.16%	35.60%	34.2658
0.65	9.72%	34.92%	12.02%	48.65%	11.99%	48.64%	5.63%	34.96%	34.4102
0.70	9.28%	34.93%	11.54%	47.40%	11.56%	47.38%	5.06%	34.69%	34.5394
0.75	10.08%	35.60%	12.12%	46.96%	12.08%	46.92%	6.05%	34.65%	34.6499
0.80	9.82%	34.53%	11.53%	45.71%	11.49%	45.71%	6.04%	33.21%	34.7484
0.85	9.61%	34.57%	10.94%	44.77%	10.94%	44.74%	5.96%	33.02%	34.8273
0.90	10.47%	34.77%	11.47%	43.99%	11.48%	43.96%	6.48%	32.98%	34.9176
0.95	10.23%	35.08%	11.15%	43.64%	11.14%	43.61%	6.42%	32.60%	34.9878
1.00	10.02%	34.44%	10.87%	42.77%	10.87%	42.75%	6.08%	32.15%	35.0498
1.05	9.89%	35.02%	10.79%	42.42%	10.79%	42.40%	5.75%	32.01%	35.1099
1.10	9.42%	33.73%	10.43%	41.29%	10.43%	41.20%	5.37%	30.17%	35.1653
1.15	10.18%	34.00%	10.58%	40.51%	10.58%	40.50%	6.44%	30.57%	35.2172
1.20	10.12%	34.70%	10.66%	41.01%	10.65%	41.02%	6.50%	31.02%	35.2588
1.25	9.44%	33.39%	9.93%	39.40%	9.94%	39.40%	5.86%	29.81%	35.3034
1.30	9.95%	35.33%	10.76%	41.24%	10.74%	41.25%	6.17%	31.11%	35.3382

Notes: (1) This table only reports the inference results for $\hat{\beta}_{n,\hat{w}_n,1}$, the averaging estimator of β_1. The results for the other thee coordinates are reported in Tables D.1 - D.3 in Appendix D.

(2) All results are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size. The two-step inference method uses $S = 1000$ replications to simulate the distribution of $\bar{\xi}_{F,d} \equiv (1 - w_F)\xi_{F,SP} + w_F(\xi_{F,P} + d)$ in (3.22).

(3) The naive inference methods treat the averaging weight $\hat{w}_n$ as non-random, and hence underestimate the randomness in $\hat{\beta}_{n,\hat{w}_n,1}$. Two naive methods are reported here: the first uses the common estimators of $V_{F,P}$ and C_F, which might be biased under misspecification (see the discussion after (C.3) for details); and the second combines $\widehat{V}_{n,P}$ and $\widehat{C}_n$ given in (C.2) and (C.3) with the robust influence function $\psi_{n,P}(z)$ in Appendix D, which are robust under misspecification (robust SE).

(4) The test value for the "Size" columns is 4, the true value of β_1; the test value for the "Power" columns is 0.

(5) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Appendix A Proofs of the Theorems

Proof of Lemma 1

Proof. Part (i). Note that $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$ are consistent estimators of $V_{F,SP}$, $V_{F,P}$ and C_F, respectively, then the result in part (i) follows by Condition 2(i) and the continuous mapping theorem.

Part (ii). Because the probability limits of $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$ are finite and $\|n^{1/2}(\hat{\beta}_{n,P}-\hat{\beta}_{n,SP})\| \xrightarrow{p.} \infty$, one has $\hat{w}_n \xrightarrow{p.} 0$ by the continuous mapping theorem. This, combined with Slutsky's theorem, implies that $n^{1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_F) \xrightarrow{d.} \xi_{F,SP}$.

■

The following notation will be used in the proofs. Let C and κ be generic symbols for positive constants that might take different values at each appearance. For any $u_{F,d} \in \mathcal{U} \cup \mathcal{U}_\infty$ (defined in (3.24) - (3.26)) and any positive finite ζ, define

$$R_\zeta(u_{F,d}) \equiv \mathbb{E}_F\left(\min\left\{\xi'_{F,SP}\Upsilon\xi_{F,SP}, \zeta\right\}\right), \tag{A.1}$$

$$\bar{R}_\zeta(u_{F,d}) \equiv \begin{cases} \mathbb{E}_F\left(\min\left\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}, \zeta\right\}\right), & \text{if } \|d\| < \infty \text{ (i.e., } u_{F,d} \in \mathcal{U}), \\ \mathbb{E}_F\left(\min\left\{\xi'_{F,SP}\Upsilon\xi_{F,SP}, \zeta\right\}\right), & \text{if } \|d\| = \infty \text{ (i.e., } u_{F,d} \in \mathcal{U}_\infty), \end{cases} \tag{A.2}$$

$$r_\zeta(u_{F,d}) \equiv \bar{R}_\zeta(u_{F,d}) - R_\zeta(u_{F,d}) \tag{A.3}$$

$$= \begin{cases} \mathbb{E}_F\left(\min\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}, \zeta\} - \min\{\xi'_{F,SP}\Upsilon\xi_{F,SP}, \zeta\}\right), & \text{if } \|d\| < \infty \text{ (i.e., } u_{F,d} \in \mathcal{U}), \\ 0, & \text{if } \|d\| = \infty \text{ (i.e., } u_{F,d} \in \mathcal{U}_\infty), \end{cases} \tag{A.4}$$

$$r(u_{F,d}) \equiv \begin{cases} \mathbb{E}_F\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} - \xi'_{F,SP}\Upsilon\xi_{F,SP}\right), & \text{if } \|d\| < \infty \text{ (i.e., } u_{F,d} \in \mathcal{U}), \\ 0, & \text{if } \|d\| = \infty \text{ (i.e., } u_{F,d} \in \mathcal{U}_\infty). \end{cases} \tag{A.5}$$

Note that $r(u_{F,d})$ in (A.5) coincides with what is defined in (3.27). For any positive finite ζ, define

$$\begin{aligned} Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) &\equiv \limsup_{n\to\infty} \overline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta) \\ &= \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{E}_F[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_F) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_F)], \end{aligned} \tag{A.6}$$

$$\begin{aligned} Asy\underline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) &\equiv \liminf_{n\to\infty} \underline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta) \\ &= \liminf_{n\to\infty} \inf_{F\in\mathcal{F}} \mathbb{E}_F[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_F) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_F)], \end{aligned} \tag{A.7}$$

where $\overline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta)$ and $\underline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta)$ are defined in (3.12) and (3.11).

The proofs of the following Lemmas A.1 to A.4 can be found in Appendix B.

Lemma A.1. *Suppose Conditions 1 - 3 hold. Then*

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \le \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}, \tag{A.8}$$

$$Asy\underline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \ge \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}. \tag{A.9}$$

Lemma A.2. *Suppose Conditions 1 - 3 hold. Then*

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \ge \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}, \tag{A.10}$$

$$Asy\underline{RD}_{\zeta}(\hat{\beta}_{n,\hat{w}_n},\hat{\beta}_{n,SP}) \leq \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}),0\right\}. \tag{A.11}$$

Lemma A.3. *Suppose: (i) Conditions 1 - 3 hold; (ii) $tr(A_F)>0$ and $tr(B_F)>0$, with A_F and B_F defined in* (3.20)*. Then*

$$\sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left(\xi'_{F,SP}\Upsilon\xi_{F,SP}\right)^2\right] \leq C, \tag{A.12}$$

$$\sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}\right)^2\right] \leq C. \tag{A.13}$$

Lemma A.4. *Suppose: (i) Conditions 1 - 3 hold; (ii) $tr(A_F)>0$ and $tr(B_F)>0$, with A_F and B_F defined in* (3.20)*. Then*

$$\lim_{\zeta\to\infty}\sup_{u_{F,d}\in\mathcal{U}} |r_{\zeta}(u_{F,d}) - r(u_{F,d})| = 0. \tag{A.14}$$

Proof of Lemma 2

Proof. First, combining Lemmas A.1 and A.2 gives

$$Asy\overline{RD}_{\zeta}(\hat{\beta}_{n,\hat{w}_n},\hat{\beta}_{n,SP}) = \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}),0\right\}, \tag{A.15}$$

$$Asy\underline{RD}_{\zeta}(\hat{\beta}_{n,\hat{w}_n},\hat{\beta}_{n,SP}) = \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}),0\right\}, \tag{A.16}$$

for any finite $\zeta>0$. Then note that Lemma A.4 implies[29]

$$\lim_{\zeta\to\infty}\sup_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}) = \sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}), \text{ and } \lim_{\zeta\to\infty}\inf_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}) = \inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d}). \tag{A.17}$$

Moreover, note that for $u_{F,d}\in\mathcal{U}_{\infty}$ (defined in (3.26)), (A.4) and (A.5) imply that $r_{\zeta}(u_{F,d}) = r(u_{F,d}) = 0$. Furthermore, since $\max\{x,0\}$ and $\min\{x,0\}$ are both continuous functions of x, the equalities in (A.17) remain valid after applying these continuous functions; that is,

$$\lim_{\zeta\to\infty}\max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}),0\right\} = \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}),0\right\}, \tag{A.18}$$

$$\lim_{\zeta\to\infty}\min\left\{\inf_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}),0\right\} = \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d}),0\right\}. \tag{A.19}$$

Combining (A.15), (A.18), and the definitions of $Asy\overline{RD}(\hat{\beta}_{n,\hat{w}_n},\hat{\beta}_{n,SP})$ in (3.14) and of $Asy\overline{RD}_{\zeta}(\hat{\beta}_{n,\hat{w}_n},\hat{\beta}_{n,SP})$ in (A.6) gives the result in (3.28). Combining (A.16), (A.19), and the definitions of $Asy\underline{RD}(\hat{\beta}_{n,\hat{w}_n},\hat{\beta}_{n,SP})$ in (3.13) and of $Asy\underline{RD}_{\zeta}(\hat{\beta}_{n,\hat{w}_n},\hat{\beta}_{n,SP})$ in (A.7) gives the result in (3.29). ■

[29] This is because Lemma A.4 means that for $\forall\epsilon>0$, there exists a large enough number C such that for all $\zeta\geq C$ one has $\sup_{u_{F,d}\in\mathcal{U}}|r_{\zeta}(u_{F,d})-r(u_{F,d})|<\epsilon$. This implies that for $\zeta\geq C$ and $\forall u_{F,d}\in\mathcal{U}$, one has $r(u_{F,d})-\epsilon<r_{\zeta}(u_{F,d})<r(u_{F,d})+\epsilon$. The two inequalities here remain holding when the supreme operator is applied on the three expressions, and note that ϵ does not vary with $u_{F,d}$, so for $\zeta\geq C$, one gets $\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d})-\epsilon<\sup_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d})<\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d})+\epsilon$. This in turn immediately implies that for $\zeta\geq C$, one has $\left|\sup_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d)}-\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d})\right|<\epsilon$; that is, $\lim_{\zeta\to\infty}\sup_{u_{F,d}\in\mathcal{U}} r_{\zeta}(u_{F,d}) = \sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d})$. Similar relationship for the infimum can be shown using the same argument.

Proof of Theorem 1

Proof. By Lemma 2, it suffices to show that $\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}) \leq 0$ and $\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d}) < 0$. By the definition of $\bar{\xi}_{F,d}$ in (3.22), one gets

$$\begin{aligned}\mathbb{E}(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}) =&\mathbb{E}(\xi'_{F,SP}\Upsilon\xi_{F,SP}) + 2\mathbb{E}[w_F(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon\xi_{F,SP}]\\ &+\mathbb{E}[w_F^2(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon(\xi_{F,P}+d-\xi_{F,SP})].\end{aligned}$$

By the definitions of w_F in (3.21) and of A_F and B_F in (3.20), this implies that for any $u_{F,d}\in\mathcal{U}$ (defined in (3.25)),

$$r(u_{F,d}) = 2\text{tr}(A_F)J_{1,F} + [\text{tr}(A_F)]^2 J_{2,F}, \tag{A.20}$$

where

$$\begin{aligned}J_{1,F} &\equiv \mathbb{E}\left[\frac{(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon\xi_{F,SP}}{\text{tr}(B_F)+(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon(\xi_{F,P}+d-\xi_{F,SP})}\right],\\ J_{2,F} &\equiv \mathbb{E}\left[\frac{(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon(\xi_{F,P}+d-\xi_{F,SP})}{[\text{tr}(B_F)+(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon(\xi_{F,P}+d-\xi_{F,SP})]^2}\right].\end{aligned}$$

Define

$$D \equiv [\ -I_k \quad I_k\ ]'\Upsilon[\ -I_k \quad I_k\ ], \quad \tilde{d}\equiv(0_{1\times k}, d')', \quad \textit{and} \quad E\equiv[\ -I_k \quad I_k\ ]'\Upsilon[\ I_k \quad 0_{k\times k}\ ], \tag{A.21}$$

then $J_{1,F}$ and $J_{2,F}$ can be re-written as

$$\begin{aligned}J_{1,F} &= \mathbb{E}\left[\frac{(\tilde{\xi}_F+\tilde{d})'E(\tilde{\xi}_F+\tilde{d})}{\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right],\\ J_{2,F} &= \mathbb{E}\left[\frac{(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}{[\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right],\end{aligned}$$

where $\tilde{\xi}_F$ is defined in Condition 2(i).

First consider bounding $J_{1,F}$. Define a function $\eta_F(x):\mathbb{R}^{2k}\mapsto\mathbb{R}^{2k}$ for any $x\in\mathbb{R}^{2k}$ as follows

$$\eta_F(x) \equiv \frac{x}{\text{tr}(B_F)+x'Dx}.$$

Its derivative (transposed) is then

$$\frac{\partial}{\partial x}\eta_F(x)' = \frac{I_{2k}}{\text{tr}(B_F)+x'Dx} - \frac{2Dxx'}{[\text{tr}(B_F)+x'Dx]^2}.$$

Note that $J_{1,F} = \mathbb{E}[\eta_F(\tilde{\xi}_F+\tilde{d})'E(\tilde{\xi}_F+\tilde{d})]$ and $\text{tr}(E\tilde{V}_F) = -\text{tr}[\Upsilon(V_{F,SP}-C_F)] = -\text{tr}(A_F)$, where $\tilde{V}_F$ is defined in Condition 2(i). Applying Lemma 2 in Hansen (2016), which is a matrix version of Stein's lemma (Stein, 1956), to $J_{1,F}$, one gets

$$\begin{aligned}J_{1,F} =&\mathbb{E}\left[\text{tr}\left(\frac{\partial}{\partial x}\eta_F(\tilde{\xi}_F+\tilde{d})'E\tilde{V}_F\right)\right]\\ =&\mathbb{E}\left[\frac{-\text{tr}(A_F)}{\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right] - 2\mathbb{E}\left[\frac{\text{tr}[D(\tilde{\xi}_F+\tilde{d})(\tilde{\xi}_F+\tilde{d})'E\tilde{V}_F]}{[\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right]\end{aligned}$$

$$=\mathbb{E}\left[\frac{-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right]+2\mathbb{E}\left[\frac{-(\tilde{\xi}_F+\tilde{d})'E\tilde{V}_FD(\tilde{\xi}_F+\tilde{d})}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right]. \tag{A.22}$$

By the definitions of A_F, D and E (in (3.20) and (A.21)), one has

$$\begin{aligned}
&-(\tilde{\xi}_F+\tilde{d})'E\tilde{V}_FD(\tilde{\xi}_F+\tilde{d})\\
=&(\tilde{\xi}_F+\tilde{d})'[\ -I_k\quad I_k\ ]'\Upsilon(V_{F,SP}-C_F)\Upsilon[\ -I_k\quad I_k\ ](\tilde{\xi}_F+\tilde{d})\\
\leq&\rho_{\max}[\Upsilon^{1/2}(V_{F,SP}-C_F)\Upsilon^{1/2}](\tilde{\xi}_F+\tilde{d})'[\ -I_k\quad I_k\ ]'\Upsilon[\ -I_k\quad I_k\ ](\tilde{\xi}_F+\tilde{d})\\
=&\rho_{\max}(A_F)(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d}),
\end{aligned} \tag{A.23}$$

where the last equality holds due to $\rho_{\max}[\Upsilon^{1/2}(V_{F,SP}-C_F)\Upsilon^{1/2}]=\rho_{\max}[\Upsilon(V_{F,SP}-C_F)]=\rho_{\max}(A_F)$. Combining the results in (A.22) and (A.23) gives

$$\begin{aligned}
J_{1,F}\leq&\mathbb{E}\left[\frac{-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right]+2\mathbb{E}\left[\frac{\rho_{\max}(A_F)(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right]\\
=&\mathbb{E}\left[\frac{2\rho_{\max}(A_F)-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right]-\mathbb{E}\left[\frac{2\rho_{\max}(A_F)\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right].
\end{aligned} \tag{A.24}$$

Next consider $J_{2,F}$. By applying some algebraic operations to $J_{2,F}$, one gets

$$\begin{aligned}
J_{2,F}=&\mathbb{E}\left[\frac{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})-\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right]\\
=&\mathbb{E}\left[\frac{1}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right]-\mathbb{E}\left[\frac{\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right].
\end{aligned} \tag{A.25}$$

Combining (A.20), (A.24) and (A.25) gives

$$\begin{aligned}
r(u_{F,d})\leq&2\mathrm{tr}(A_F)\mathbb{E}\left[\frac{2\rho_{\max}(A_F)-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right]\\
&-2\mathrm{tr}(A_F)\mathbb{E}\left[\frac{2\rho_{\max}(A_F)\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right]\\
&+[\mathrm{tr}(A_F)]^2\mathbb{E}\left[\frac{1}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right]\\
&-[\mathrm{tr}(A_F)]^2\mathbb{E}\left[\frac{\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right]\\
=&\mathbb{E}\left[\frac{\mathrm{tr}(A_F)[4\rho_{\max}(A_F)-\mathrm{tr}(A_F)]}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right]-\mathbb{E}\left[\frac{\mathrm{tr}(A_F)\mathrm{tr}(B_F)[4\rho_{\max}(A_F)+\mathrm{tr}(A_F)]}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right].
\end{aligned} \tag{A.26}$$

If $\mathrm{tr}(A_F)\geq 0$ and $\mathrm{tr}(B_F)\geq 0$, then $\rho_{\max}(A_F)\geq 0$, and then the second term in (A.26) will be non-positive. If, in addition, $\mathrm{tr}(A_F)\geq 4\rho_{\max}(A_F)$, then the first term in (A.26) will be non-positive. Together they imply $r(u_{F,d})\leq 0$ for any $u_{F,d}\in\mathcal{U}$, which in turn implies $\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d})\leq 0$. So, (3.16) holds in consequence.

If, furthermore, $\mathrm{tr}(B_F)>0$ for some $F\in\mathcal{F}$, then $\mathrm{tr}(A_F)>0$ and $\rho_{\max}(A_F)>0$, and then the second term in (A.26) will be strictly negative. This implies $r(u_{F,d})<0$ for some $u_{F,d}\in\mathcal{U}$, which in turn implies $\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d})<0$. So, (3.15) holds in consequence.

Note that the proof here relies on Lemma 2, which requires $\mathrm{tr}(A_F)>0$ and $\mathrm{tr}(B_F)>0$ as premises, so

the effective conditions are those stated in the theorem. ■

Proof of Lemma 3

Proof. For any given d, one has

$$\begin{aligned}
1-\alpha_2 = & \lim_{n\to\infty} P_F\left(\sqrt{n}(\hat{\beta}_{n,\hat{w}_n}-\beta_F)\in CI_{1-\alpha_2}(\Delta_n|d,\widehat{V})\right) \\
\leq & \lim_{n\to\infty} P_F\left(\sqrt{n}(\hat{\beta}_{n,\hat{w}_n}-\beta_F)\in CI_{1-\alpha_2}(\Delta_n|d,\widehat{V}),\ d\in CI_{1-\alpha_1}(d|\hat{\beta},\widehat{V})\right) \\
& + \lim_{n\to\infty} P_F\left(d\notin CI_{1-\alpha_1}(d|\hat{\beta},\widehat{V})\right) \\
\leq & \lim_{n\to\infty} P_F\left(\beta_F\in CI_{1-\alpha}(\beta|\hat{\beta},\widehat{V})\right)+\alpha_1,
\end{aligned}$$

where the first equality holds by the way in which $CI_{1-\alpha_2}(\Delta_n|d,\widehat{V})$ is constructed, the last inequality holds by the definitions of $CI_{1-\alpha}(\beta|\hat{\beta},\widehat{V})$ in (3.31) and of $CI_{1-\alpha_1}(d|\hat{\beta},\widehat{V})$ in (3.30). The last inequality in turn immediately implies the validity of (3.32) for $\alpha_1+\alpha_2=\alpha$. ■

References

Ackerberg, D., X. Chen, and J. Hahn (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics 94*(2), 481–498.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014). Asymptotic efficiency of semiparametric two-step GMM. *Review of Economic Studies 81*(3), 919–943.

Ahn, H., H. Ichimura, and J. L. Powell (1996). Simple estimators for monotone index models. *Manuscript, Department of Economics, UC Berkeley*.

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1-2), 3–29.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy 113*(1), 151–184.

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica 62*(1), 43–72.

Andrews, D. W., X. Cheng, and P. Guggenberger (2011). Generic results for establishing the asymptotic size of confidence sets and tests. *Manuscript*.

Andrews, D. W. and P. Guggenberger (2009). Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory 25*(3), 669–709.

Andrews, D. W. and P. Guggenberger (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory 26*(2), 426–468.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017). Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics 132*(4), 1553–1592.

Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics 12*(1), 77–108.

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

Bickel, P. J., C. A. Klaassen, J. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

Bickel, P. J. and Y. Ritov (2003). Nonparametric estimators which can be "plugged-in". *Annals of Statistics 31*(4), 1033–1053.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica 58*(6), 1443–1458.

Blundell, R. and J. L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics*. Cambridge University Press.

Blundell, R. W. and J. L. Powell (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies 71*(3), 655–679.

Bonhomme, S. and M. Weidner (2018). Minimizing sensitivity to model misspecification. *arXiv preprint, arXiv:1807.02161*.

Buchholz, N., M. Shum, and H. Xu (2021). Semiparametric estimation of dynamic discrete choice models. *Journal of Econometrics 223*(2), 312–327.

Cao, W., A. A. Tsiatis, and M. Davidian (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*(3), 723–734.

Chen, X., O. Linton, and I. Van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica 71*(5), 1591–1608.

Cheng, X., Z. Liao, and R. Shi (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics 10*(3), 931–979.

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

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica 90*(4), 1501–1535.

Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge University Press.

Crepon, B., F. Kramarz, and A. Trognon (1997). Parameters of interest, nuisance parameters and orthogonality conditions. An application to autoregressive error component models. *Journal of Econometrics 82*(1), 135–156.

DiTraglia, F. J. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics 195*(2), 187–208.

Donald, S. G. and W. K. Newey (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis 50*(1), 30–40.

Fan, Y. and A. Ullah (1999). Asymptotic normality of a combined regression estimator. *Journal of Multivariate Analysis 71*(2), 191–240.

Fessler, P. and M. Kasy (2019). How to use economic theory to improve estimators: Shrinking toward theoretical restrictions. *Review of Economics and Statistics 101*(4), 681–698.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), 259–276.

Fourdrinier, D., W. E. Strawderman, and M. T. Wells (2018). *Shrinkage estimation*. Springer.

Gallant, A. R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica 55*(2), 363–390.

Hahn, J. and Z. Liao (2021). Bootstrap standard error estimates and inference. *Econometrica 89*(4), 1963–1977.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics 35*(2-3), 303–316.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics 5*(3), 495–530.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics 190*(1), 115–132.

Hansen, B. E. (2017). Stein-like 2SLS estimator. *Econometric Reviews 36*(6-9), 840–852.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement*, Volume 5, pp. 475–492. NBER.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 153–161.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Hjort, N. L. and G. Claeskens (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association 101*(476), 1449–1464.

Honoré, B. E. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica*, 533–565.

Hotz, V. J. and R. A. Miller (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies 60*(3), 497–529.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric M-estimators. *Journal of Econometrics 159*(2), 252–266.

Ichimura, H. and W. Newey (2017). The influence function of semiparametric estimators. *Working paper*.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review 93*(2), 126–132.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 361–379. University of California Press.

Judge, G. G. and R. C. Mittelhammer (2004). A semiparametric basis for combining estimation problems under quadratic loss. *Journal of the American Statistical Association 99*(466), 479–487.

Judge, G. G. and R. C. Mittelhammer (2007). Estimation and inference in the case of competing sets of estimating equations. *Journal of Econometrics 138*(2), 513–531.

Keane, M. P. and K. I. Wolpin (1997). The career decisions of young men. *Journal of Political Economy 105*(3), 473–522.

Kitagawa, T. and C. Muris (2016). Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics 193*(1), 271–289.

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 387–421.

Le Cam, L. (1972). Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 245–261. University of California Press.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review 75*(3), 308–313.

Lee, L.-F. (1982). Some approaches to the correction of selectivity bias. *The Review of Economic Studies 49*(3), 355–372.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics 142*(1), 201–211.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics 186*(1), 142–159.

Lu, X. and L. Su (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics 188*(1), 40–58.

Magnus, J. R., O. Powell, and P. Prüfer (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics 154*(2), 139–153.

Mittelhammer, R. C. and G. G. Judge (2005). Combining estimators to improve structural model estimation and inference under quadratic loss. *Journal of Econometrics 128*(1), 1–29.

Mukhin, Y. (2018). Sensitivity of regular estimators. *arXiv preprint, arXiv:1805.08883*.

Nelson, F. D. (1984). Efficiency of the two-step estimator for models with endogenous sample selection. *Journal of Econometrics 24*, 181–196.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics 5*(2), 99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 1349–1382.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal 12*, S217–S229.

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

Newey, W. K. and J. L. Powell (1993). Efficiency bounds for some semiparametric selection models. *Journal of Econometrics 58*(1-2), 169–184.

Newey, W. K. and J. L. Powell (1999). Two-step estimation, optimal moment conditions, and sample selection models. *MIT working papers*.

Newey, W. K., J. L. Powell, and J. R. Walker (1990). Semiparametric estimation of selection models: some empirical results. *The American Economic Review 80*(2), 324–328.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and Statsitics*, 213–234.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics 37*(2), 187–204.

Pakes, A. and S. Olley (1995). A limit theorem for a smooth class of semiparametric estimators. *Journal of Econometrics 65*(1), 295–332.

Peng, J. and Y. Yang (2021). On improvability of model selection by model averaging. *Journal of Econometrics*.

Powell, J. L. (1986). Symmetrically trimmed least squares estimation for Tobit models. *Econometrica*, 1435–1460.

Powell, J. L. (1994). Estimation of semiparametric models. *Handbook of Econometrics 4*, 2443–2521.

Powell, J. L. (2001). Semiparametric estimation of censored selection models. In C. Hsiao, K. Morimune, and J. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, Volume 13, pp. 165–196.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Robinson, P. M. (1989). Hypothesis testing in semiparametric and nonparametric models for econometric time series. *The Review of Economic Studies 56*(4), 511–534.

Rosenbaum, P. R. and D. B. Rubin (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological) 45*(2), 212–218.

Rubin, D. B. and M. J. van der Laan (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics 4*(1), Article 5.

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

Shao, J. (1992). Bootstrap variance estimators with truncation. *Statistics & Probability Letters 15*(2), 95–101.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, 123–137.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Contributions to the Theory of Statistics. University of California Press.

Tsiatis, A. A., M. Davidian, and W. Cao (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics 67*(2), 536–545.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Wales, T. J. and A. D. Woodland (1980, June). Sample selectivity and the estimation of labor supply functions. *International Economic Review 21*(2), 437–468.

Wan, A. T., X. Zhang, and G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics 156*(2), 277–283.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association 96*(454), 574–588.

Yang, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statistica Sinica 13*(3), 783–809.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika 92*(4), 937–950.

Zhang, X. and H. Liang (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics 39*(1), 174–200.

Supplementary Material on "*An Averaging Estimator for Two Step M Estimation in Semiparametric Models*"

Ruoyao Shi[1]

September, 2022

Appendix B Proofs of the Lemmas in Appendix A

Proof of Lemma A.1

Proof. The following proves inequality (A.8). By the definition of supremum and the definition of $Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP})$ in (A.6), there exists a sequence of DGPs, denoted by $\{F_n\}_{n\in\mathbb{N}}$, such that

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \limsup_{n\to\infty} \mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})].$$

The real sequence $\{\mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})]\}_{n\in\mathbb{N}}$ itself may not be convergent, but by the definition of limsup, there exists a subsequence of $\{n\}_{n\in\mathbb{N}}$, denoted by $\{p_n\}_{n\in\mathbb{N}}$, such that the corresponding real subsequence $\{\mathbb{E}_{F_{p_n}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})]\}_{n\in\mathbb{N}}$ is convergent, where $\{F_{p_n}\}_{n\in\mathbb{N}}$ denotes the subsequence of DGPs corresponding to $\{p_n\}_{n\in\mathbb{N}}$, then

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \lim_{n\to\infty} \mathbb{E}_{F_{p_n}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})]. \tag{B.1}$$

Now consider the sequence of k-dimensional vectors $\{p_n^{1/2}\delta_{F_{p_n}}\}_{n\in\mathbb{N}}$, and let $\{p_n^{1/2}\delta_{F_{p_n},\iota}\}_{n\in\mathbb{N}}$ $(\iota = 1, \ldots, k)$ denote their ιth coordinates. For $\iota = 1$, one has either (i) $\limsup_{n\to\infty} \left|p_n^{1/2}\delta_{F_{p_n},\iota}\right| < \infty$, or (ii) $\limsup_{n\to\infty} \left|p_n^{1/2}\delta_{F_{p_n},\iota}\right| = \infty$. For case (i), there exists some subsequence $\{p_{n,\iota}\}_{n\in\mathbb{N}}$ such that $p_{n,\iota}^{1/2}\delta_{F_{p_{n,\iota}},\iota} \to d_\iota$ for some $d_\iota \in \mathbb{R}$, by the definition of limsup. For case (ii), there exists some subsequence $\{p_{n,\iota}\}_{n\in\mathbb{N}}$ such that $p_{n,\iota}^{1/2}\delta_{F_{p_{n,\iota}},\iota} \to \infty$ or $-\infty$, by the definition of limsup. In both cases, therefore, there exists some subsequence $\{p_{n,\iota}\}_{n\in\mathbb{N}}$ such that $p_{n,\iota}^{1/2}\delta_{F_{p_{n,\iota}},\iota} \to d_\iota$ for some $d_\iota \in \mathbb{R}_\infty$. Since k is finite, one can sequentially apply the same argument to all coordinates $\iota = 2, \ldots, k$ and let the resulting subsequence be denoted by $\{p_{n,k}\}_{n\in\mathbb{N}}$, which satisfies $p_{n,k}^{1/2}\delta_{F_{p_{n,k}}} \to d$ for some $d \in \mathbb{R}_\infty^k$. Next consider $\{S(F_{p_{n,k}})\}_{n\in\mathbb{N}}$, the sequence of nuisance parameter vectors in $\mathcal{S}$ induced by the DGPs $\{F_{p_{n,k}}\}_{n\in\mathbb{N}}$. $\{S(F_{p_{n,k}})\}_{n\in\mathbb{N}}$ itself may not be convergent, but since $\mathcal{S}$ is compact by Condition 3(i), then there exists a convergent subsequence, denoted by $\{S(F_{p_n^*})\}_{n\in\mathbb{N}}$, such that $S(F_{p_n^*}) \longrightarrow s^*$ with some $s^* \in \mathcal{S}$. Moreover, by Condition 3(ii), there exists a DGP F^* in $\mathcal{F}$ such that $S(F^*) = s^*$. This shows that there exists some subsequence $\{p_n^*\}_{n\in\mathbb{N}}$ of $\{p_n\}_{n\in\mathbb{N}}$ such that

$$p_n^{*1/2}\delta_{F_{p_n^*}} \to d \text{ for some } d \in \mathbb{R}_\infty \text{ and } S(F_{p_n^*}) \longrightarrow S(F^*) \text{ for some } F^* \in \mathcal{F}. \tag{B.2}$$

Note that for any subsequence of $\{p_n\}_{n\in\mathbb{N}}$, the limit of the right hand side in (B.1) remains the same, which implies

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \lim_{n\to\infty} \mathbb{E}_{F_{p_n^*}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n^*}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n^*}})]. \tag{B.3}$$

The definitions of $\ell(\hat{\beta}_n, \beta)$ in (2.1) and of $\ell_\zeta(\hat{\beta}_n, \beta)$ in (3.10), as well as (A.3) suggest that in order to prove (A.8), one needs to link the right hand side of (B.3) with $R_\zeta(u_{F,d})$ and $\bar{R}_\zeta(u_{F,d})$ defined in (A.1) and

[1]Ruoyao Shi: Department of Economics, UC Riverside, USA. Email: ruoyao.shi@ucr.edu.

(A.2). First consider the case where $\|d\| < \infty$ in (B.2). By Condition 2(i) and Lemma 1(i),

$$p_n^{*1/2}(\hat{\beta}_{n,SP} - \beta_{F_{p_n^*}}) \xrightarrow{d.} \xi_{F,SP} \text{ and } p_n^{*1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_{F_{p_n^*}}) \xrightarrow{d.} \bar{\xi}_{F,d},$$

which combined with the continuous mapping theorem implies that

$$\ell(\hat{\beta}_{n,SP}, \beta_{F_{p_n^*}}) \xrightarrow{d.} \xi'_{F,SP}\Upsilon\xi_{F,SP} \text{ and } \ell(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n^*}}) \xrightarrow{d.} \bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}.$$

Since Υ is positive semi-definite, $\xi'_{F,SP}\Upsilon\xi_{F,SP}$ and $\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}$ are both nonnegative. Note that the function $f(x) \equiv \min\{x, \zeta\}$ is a bounded continuous function of $x \geq 0$ for fixed positive ζ. Applying the Portmanteau lemma (e.g., Lemma 2.2 in Van der Vaart, 2000) and invoking (A.1) and (A.2), one gets

$$\mathbb{E}_{F_{p_n^*}}\left[\ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n^*}})\right] \to R_\zeta(u_{F^*,d}) \text{ and } \mathbb{E}_{F_{p_n^*}}\left[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n^*}})\right] \to \bar{R}_\zeta(u_{F^*,d}). \tag{B.4}$$

Next consider the case where $\|d\| = \infty$ in (B.2). By Condition 2(ii) and Lemma 1(ii),

$$p_n^{*1/2}(\hat{\beta}_{n,SP} - \beta_{F_{p_n^*}}) \xrightarrow{d.} \xi_{F,SP} \text{ and } p_n^{*1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_{F_{p_n^*}}) \xrightarrow{d.} \xi_{F,SP}.$$

Using the same argument, one also gets (B.4) in this case. Combining (A.3), (B.3) and (B.4), one can unify the two cases and write

$$\begin{aligned} Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) &= r_\zeta(u_{F^*,d}), \text{ for some } F^* \in \mathcal{F} \text{ and some } d \in \mathbb{R}^k_\infty \\ &\leq \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), \sup_{u_{F,d}\in\mathcal{U}_\infty} r_\zeta(u_{F,d})\right\} \\ &= \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}. \end{aligned}$$

This proves (A.8).

The proof of (A.9) follows the same argument and hence is omitted here. ■

Proof of Lemma A.2

Proof. The following proves inequality (A.10). By the definition of $\mathcal{U}$ in (3.25), $\|d\| < \infty$ and $\delta_F = 0$ for any $F \in \mathcal{F}$ such that $u_{F,d} \in \mathcal{U}$. For any $u_{F,d} \in \mathcal{U}$, let N_{ϵ_F} denote the smallest n such that $n^{-1/2}\|d\| < \epsilon_F$, where ϵ_F satisfies Condition 3(ii). Then by Condition 3(ii), for each $n \geq N_{\epsilon_F}$, there is an $F_n \in \mathcal{F}$ with $\delta_{F_n} = n^{-1/2}d$ and $\|\bar{S}(F_n) - \bar{S}(F)\| \leq n^{-\kappa/2}C\|d\|^\kappa$ for some $C, \kappa > 0$. For each $n < N_{\epsilon_F}$, let $F_n = F$. Thus, a sequence of DGPs $\{F_n\}_{n\in\mathbb{N}}$ in $\mathcal{F}$ satisfying $n^{1/2}\delta_{F_n} \to d$ and $\bar{S}(F_n) \to \bar{S}(F)$ is constructed for any $u_{F,d} \in \mathcal{U}$. Recalling the definition of $\bar{S}(F)$ in (3.17), this immediately implies that for such $\{F_n\}_{n\in\mathbb{N}}$,

$$n^{1/2}\delta_{F_n} \to d \in \mathbb{R}^k,\ V_{F_n,SP} \to V_{F,SP},\ C_{F_n} \to C_F, \text{ and } V_{F_n,P} \to V_{F,P}. \tag{B.5}$$

The real sequence $\{\mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})]\}_{n\in\mathbb{N}}$ that corresponds to $\{F_n\}_{n\in\mathbb{N}}$ may not be convergent, but by the definition of lim sup, there exists a subsequence $\{p_n\}_{n\in\mathbb{N}}$ of $\{n\}_{n\in\mathbb{N}}$ such that the corresponding real sequence $\{\mathbb{E}_{F_{p_n}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})]\}_{n\in\mathbb{N}}$ is convergent and

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})] = \limsup_{n\to\infty} \mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})]. \tag{B.6}$$

Since $\{p_n\}_{n\in\mathbb{N}}$ is a subsequence of $\{n\}_{n\in\mathbb{N}}$, (B.5) implies that

$$n^{1/2}\delta_{F_{p_n}} \to d \in \mathbb{R}^k,\ V_{F_{p_n},SP} \to V_{F,SP},\ C_{F_n} \to C_F, \text{ and } V_{F_{p_n},P} \to V_{F,P}. \tag{B.7}$$

Combined with Condition 2(i) and Lemma 1(i), this implies that

$$p_n^{1/2}(\hat{\beta}_{n,SP} - \beta_{F_{p_n}}) \xrightarrow{d.} \xi_{F,SP}, \text{ and } p_n^{1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_{F_{p_n}}) \xrightarrow{d.} \bar{\xi}_{F,d},$$

which, combined with the continuous mapping theorem, in turn implies that

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\left[\ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})\right] = R_\zeta(u_{F,d}), \text{ and } \lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\left[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}})\right] = \bar{R}_\zeta(u_{F,d}). \tag{B.8}$$

This, combined with (B.6), the definition of $Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP})$ in (A.6), the definition of supremum and the definition of $r(u_{F,d})$ in (A.3), implies that for any $u_{F,d} \in \mathcal{U}$,

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \geq \limsup_{n\to\infty} \mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})] = r(u_{F,d}),$$

which further implies that

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \geq \sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}). \tag{B.9}$$

On the other hand, by the definition of $\mathcal{U}_\infty$ in (3.26), for any $u_{F,d} \in \mathcal{U}_\infty$, $\|d\| = \infty$ and either (i) $\delta_F = 0$ or (ii) $\|\delta_F\| > 0$. For case (i), let ℓ_k be a $k \times 1$ vector of ones and let N_{ϵ_F} denote the smallest n such that $n^{-1/4}\|\ell_k\|^{1/2} = n^{-1/4}k^{1/2} < \epsilon_F$, where ϵ_F satisfies Condition 3(ii). Then by Condition 3(ii), for each $n \geq N_{\epsilon_F}$, there is an $F_n \in \mathcal{F}$ with $\delta_{F_n} = n^{-1/4}\ell_k$ and $\|\bar{S}(F_n) - \bar{S}(F)\| \leq Cn^{-\kappa/4}k^{\kappa/2}$ for some $C, \kappa > 0$. For each $n < N_{\epsilon_F}$, let $F_n = F$. For case (ii), let $F_n = F$ for all n. Thus, a sequence of DGPs $\{F_n\}_{n\in\mathbb{N}}$ in $\mathcal{F}$ satisfying $n^{1/2}\delta_{F_n} \to \infty$, $\delta_{F_n} \to F$ and $\bar{S}(F_n) \to \bar{S}(F)$ is constructed for any $u_{F,d} \in \mathcal{U}_\infty$, regardless of whether $\delta_F = 0$ or $\|\delta_F\| > 0$. Recalling the definition of $\bar{S}(F)$ in (3.17), this immediately implies that for such $\{F_n\}_{n\in\mathbb{N}}$,

$$\|n^{1/2}\delta_{F_n}\| \to \infty,\ V_{F_n,SP} \to V_{F,SP},\ C_{F_n} \to C_F, \text{ and } V_{F_n,P} \to V_{F,P}.$$

Then a similar argument used to show (B.6) - (B.8) can be applied to show that there exists a subsequence $\{p_n\}_{n\in\mathbb{N}}$ of $\{n\}_{n\in\mathbb{N}}$ such that (B.6) and (B.8) are satisfied, with the help of Condition 2(ii) and Lemma 1(ii). Combining this with the definition of $Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP})$ in (A.6), the definition of supremum and the definition of $r(u_{F,d})$ in (A.3), one gets that for any $u_{F,d} \in \mathcal{U}_\infty$,

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \geq \limsup_{n\to\infty} \mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})] = 0. \tag{B.10}$$

(A.10) immediately follows inequalities (B.9) and (B.10).

The proof of (A.11) follows the same argument and hence is omitted here. ■

Proof of Lemma A.3

Proof. For any $F \in \mathcal{F}$, since $\xi_{F,SP} \sim \mathcal{N}(0_{k\times 1}, V_{F,SP})$ by Condition 2, one gets

$$\xi'_{F,SP}\Upsilon\xi_{F,SP} \overset{d.}{=} \mathcal{Z}'V_{F,SP}^{1/2}\Upsilon V_{F,SP}^{1/2}\mathcal{Z},$$

where $\mathcal{Z} \sim \mathcal{N}(0_{k\times 1}, I_{k\times k})$. By Condition 3(i), and because Υ is a fixed matrix, there exists some constant C such that

$$\sup_{F\in\mathcal{F}} \rho_{\max}\left(V_{F,SP}^{1/2}\Upsilon V_{F,SP}^{1/2}\right) \leq C.$$

This implies that

$$\sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left(\xi_{F,SP}'\Upsilon\xi_{F,SP}\right)^2\right] \leq \sup_{u_{F,d}\in\mathcal{U}} \rho_{\max}^2\left(V_{F,SP}^{1/2}\Upsilon V_{F,SP}^{1/2}\right)\cdot \mathbb{E}[(\mathcal{Z}'\mathcal{Z})^2] \leq C,$$

where the second inequality holds because $\mathcal{Z} \sim \mathcal{N}(0_{k\times 1}, I_{k\times k})$ and that $V_{F,SP}$ does not depend on d. This proves (A.12).

By the definition of $\bar{\xi}_{F,d}$ in (3.22) and that of $\tilde{\xi}_F$ in Condition 2(i), the Cauchy-Schwarz inequality and the simple inequality $2|ab| \leq a^2 + b^2$ for any real numbers a and b, one gets

$$\begin{aligned}\bar{\xi}_{F,d}'\Upsilon\bar{\xi}_{F,d} &\leq 2\xi_{F,SP}'\Upsilon\xi_{F,SP} + 2w_F^2\left(\xi_{F,P}+d-\xi_{F,SP}\right)'\Upsilon\left(\xi_{F,P}+d-\xi_{F,SP}\right)\\ &= 2\xi_{F,SP}'\Upsilon\xi_{F,SP} + 2w_F^2\left(\tilde{\xi}_F+\tilde{d}\right)'D\left(\tilde{\xi}_F+\tilde{d}\right),\end{aligned} \tag{B.11}$$

where D and $\tilde{d}$ are defined in (A.21). Combining (B.11) and the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$ for any real numbers a and b, one gets

$$\begin{aligned}\left(\bar{\xi}_{F,d}'\Upsilon\bar{\xi}_{F,d}\right)^2 \leq& 8\left(\xi_{F,SP}'\Upsilon\xi_{F,SP}\right)^2 + 8\left[w_F^2\left(\tilde{\xi}_F+\tilde{d}\right)'D\left(\tilde{\xi}_F+\tilde{d}\right)\right]^2\\ \leq& C + 8\left[w_F^2\left(\tilde{\xi}_F+\tilde{d}\right)'D\left(\tilde{\xi}_F+\tilde{d}\right)\right]^2,\end{aligned} \tag{B.12}$$

where the second inequality is by (A.12). By the definitions of w_F in (3.21) and that of A_F and B_F in (3.20), one has

$$w_F^2\left(\tilde{\xi}_F+\tilde{d}\right)'D\left(\tilde{\xi}_F+\tilde{d}\right) = \frac{[\mathrm{tr}(A_F)]^2\left(\tilde{\xi}_F+\tilde{d}\right)'D\left(\tilde{\xi}_F+\tilde{d}\right)}{\left[\mathrm{tr}(B_F)+\left(\tilde{\xi}_F+\tilde{d}\right)'D\left(\tilde{\xi}_F+\tilde{d}\right)\right]^2} \leq C\mathrm{tr}(A_F) = C\mathrm{tr}(\Upsilon V_{F,SP}) - C\mathrm{tr}(\Upsilon C_F),$$

where the inequality follows by $\mathrm{tr}(A_F) > 0$, $\mathrm{tr}(B_F) > 0$ and that Υ being positive semi-definite implies $(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d}) \geq 0$. Combined with the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$, this implies that

$$\begin{aligned}\mathbb{E}\left[w_F^2\left(\tilde{\xi}_F+\tilde{d}\right)'D\left(\tilde{\xi}_F+\tilde{d}\right)\right]^2 &\leq 2C[\mathrm{tr}(\Upsilon V_{F,SP})]^2 + 2C[\mathrm{tr}(\Upsilon C_F)]^2\\ &\leq 2C[\mathrm{tr}(\Upsilon V_{F,SP})]^2 + 2C[\mathrm{tr}(\Upsilon V_{F,SP})]^2 \leq C,\end{aligned} \tag{B.13}$$

where the second inequality holds by Condition 2(i), Condition 3(i) and that the Cauchy-Schwarz inequality implies $C_F \leq \max\{V_{F,SP}, V_{F,P}\}$ for any $F \in \mathcal{F}$. Together, (B.12) and (B.13) imply (A.13), since the upper bound does not depend on F. ■

Proof of Lemma A.4

Proof. First note that

$$\sup_{u_{F,d}\in\mathcal{U}} \left|\mathbb{E}\left[\min\{\bar{\xi}_{F,d}'\Upsilon\bar{\xi}_{F,d}, \zeta\} - \bar{\xi}_{F,d}'\Upsilon\bar{\xi}_{F,d}\right]\right|$$

$$
\begin{aligned}
&= \sup_{u_{F,d}\in\mathcal{U}} \left|\mathbb{E}\left[\left(\zeta - \bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}\right)\mathbb{I}\left\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} > \zeta\right\}\right]\right| \\
&\leq \sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left|\zeta - \bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}\right| \cdot \mathbb{I}\left\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} > \zeta\right\}\right] \\
&\leq \zeta \sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\mathbb{I}\left\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} > \zeta\right\}\right] + \sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}\right)\cdot\mathbb{I}\left\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} > \zeta\right\}\right] \\
&\leq 2\zeta^{-1} \sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}\right)^2\right] \leq 2C\zeta^{-1},
\end{aligned} \tag{B.14}
$$

where the first equality is by the fact that $\min\{x,\zeta\} - x = (\zeta - x)\cdot\mathbb{I}\{x > \zeta\}$; the first inequality is by Jensen's inequality and the fact that an indicator function is always non-negative; the second inequality holds because $\zeta > 0$, $\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} \geq 0$, and the simple inequality $|a-b| \leq a+b$ for any non-negative real numbers a and b; the third inequality holds by Markov's inequality;[2] the fourth inequality is by (A.13) in Lemma A.3.

By (A.12) in Lemma A.3 and the same argument, one can show that

$$
\sup_{u_{F,d}\in\mathcal{U}} \left|\mathbb{E}\left[\min\{\xi_{F,SP}\Upsilon\xi_{F,SP},\zeta\} - \xi_{F,SP}\Upsilon\xi_{F,SP}\right]\right| \leq 2C\zeta^{-1}. \tag{B.15}
$$

Combining inequalities (B.14) and (B.15), the definitions of $r_\zeta(u_{F,d})$ and $r(u_{F,d})$ in (A.4) and (A.5), and the triangular inequality, one gets $\sup_{u_{F,d}\in\mathcal{U}} |r_\zeta(u_{F,d}) - r(u_{F,d})| \leq 4C\zeta^{-1}$, which immediately implies (A.14). ■

Appendix C Details on Section 2

The reasons for $\hat{w}_n \in [0,1]$ with probability one. Note that $\widehat{V}_{n,SP} + \widehat{V}_{n,P} - \widehat{C}_n - \widehat{C}'_n$ is the sample asymptotic variance-covariance matrix of $\hat{\beta}_{n,P} - \hat{\beta}_{n,SP}$ and recall that Υ is symmetric positive semi-definite, so the first term in the denominator of (2.2) is positive with probability one; the second term is a quadratic form with positive semi-definite Υ, so the denominator of (2.2) is positive with probability one. Moreover, if the parametric restrictions are correctly specified or mildly misspecified, then $V_{F,SP} \geq V_{F,P}$ (Condition 2) implies $\widehat{V}_{n,SP} \geq \widehat{V}_{n,P}$ with probability one, which, together with the Cauchy-Schwarz inequality, further implies $\widehat{V}_{n,SP} \geq \widehat{C}_n$.[3] Furthermore, if the parametric restrictions are severely misspecified, then $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$ having finite probability limits (Condition 2) implies that the second term in the denominator approaches the infinity while the other terms are finite. Together, these imply that the averaging weight $\hat{w}_n \in [0,1]$ with probability one.

Computing asymptotic variance-covariance matrices via robust influence functions. Let $\psi_{F,SP}(z)$ denote the non-centered influence function of $\hat{\beta}_{n,SP}$, let $\psi_{F,P}(z)$ denote that of $\hat{\beta}_{n,P}$, and let $\psi_{n,SP}(z)$ and $\psi_{n,P}(z)$ denote their sample analogs, respectively. Then

$$
\widehat{V}_{n,SP} = \frac{1}{n}\sum_{i=1}^n \psi_{n,SP}(Z_i)\psi'_{n,SP}(Z_i) - \left[\frac{1}{n}\sum_{i=1}^n \psi_{n,SP}(Z_i)\right]\cdot\left[\frac{1}{n}\sum_{i=1}^n \psi_{n,SP}(Z_i)\right]', \tag{C.1}
$$

$$
\widehat{V}_{n,P} = \frac{1}{n}\sum_{i=1}^n \psi_{n,P}(Z_i)\psi'_{n,P}(Z_i) - \left[\frac{1}{n}\sum_{i=1}^n \psi_{n,P}(Z_i)\right]\cdot\left[\frac{1}{n}\sum_{i=1}^n \psi_{n,P}(Z_i)\right]', \tag{C.2}
$$

[2] The first term is bounded using Chebyshev's inequality. Using the same argument as Markov's inequality, one can show that for non-negative random variable X and constant $a > 0$, $\mathbb{E}[X\cdot\mathbb{I}\{X > a\}] \leq \mathbb{E}(X^2)/a$, since $\mathbb{E}(X^2) = \mathbb{E}[X^2\cdot\mathbb{I}\{X > a\}] + \mathbb{E}[X^2\cdot\mathbb{I}\{X \leq a\}] \geq \mathbb{E}[X^2\cdot\mathbb{I}\{X > a\}] \geq a\mathbb{E}[X\cdot\mathbb{I}\{X > a\}]$. Applying this result to the second term gives the desired inequality.

[3] Condition 2(i) postulates $V_{F,SP} \geq V_{F,P}$, which is the case where the averaging is meaningful, otherwise $\hat{\beta}_{n,P}$ dominates $\hat{\beta}_{n,SP}$. Allowing for $V_{F,SP} < V_{F,P}$ is also easy and discussed in Remark 2.

$$\widehat{C}_n = \frac{1}{n}\sum_{i=1}^{n} \psi_{n,SP}(Z_i)\psi'_{n,P}(Z_i) - \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,SP}(Z_i)\right] \cdot \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,P}(Z_i)\right]'. \tag{C.3}$$

The asymptotic variance-covariance matrix estimates in (C.1) - (C.3) can then be plugged into (2.2) to compute the averaging weight.

It is worth emphasizing that the influence functions need to be valid under *potential misspecification* (e.g., Ichimura and Lee, 2010), such that the estimators $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$ and $\widehat{C}_n$ are consistent *regardless* of whether the parametric restrictions hold or not. In other words, they must be robust against misspecification of the parametric restrictions; otherwise the resulting averaging estimator might not conform to the asymptotic theory in Section 3. Appendix D will illustrate this point using the partially linear model in Section 4.

Appendix D Details on Section 4

Robust influence function in partially linear models. In this example, the estimators are based on the objective function $Q(z,\beta,h) \equiv \frac{1}{2}[y - h_1(x_2) - (x_1 - h_2(x_2))'\beta]^2$ where z represents the vector of all observed variables, so that $Q_F(\beta,h)$ in (3.1) equals to $\mathbb{E}_F[Q(Z,\beta,h)]$, where the expectation is taken with regard to the distribution F of the data Z. Under DGP F, let $h_{1F}(s) \equiv \mathbb{E}_F(Y|s(X_1,X_2)=s)$ and $h_{2F} \equiv \mathbb{E}_F(X_1|s(X_1,X_2)=s)$ denote the conditional mean functions of Y_i and X_{1i} given $s(X_{1i},X_{2i})=s$.[4] Since these functions do not depend on β, the general formula of the influence function of an estimator $\hat{\beta}_n$ robust to potential misspecification of h can be derived using Theorem 3.3 of Ichimura and Lee (2010). Borrowing their notation, one gets

$$\begin{aligned}
\Delta_1(z) \equiv D_\beta Q(z,\beta,h) &= -[y - h_1(x_2) - (x_1 - h_2(x_2))'\beta](x_1 - h_2(x_2)),\\
D_{\beta\beta'}Q(z,\beta,h) &= (x_1 - h_2(x_2))(x_1 - h_2(x_2))',\\
V_0 = \frac{d^2Q(\beta,h)}{d\beta d\beta'} &= D_{\beta\beta'}Q(\beta,h) = \mathbb{E}[(X_1 - h_2(X_2))(X_1 - h_2(X_2))'],\\
D_hQ(z,\beta,h_F)[h] &= -[y - h_{1F}(x_2) - (x_1 - h_{2F}(x_2))'\beta](h_1(x_2) - h_2(x_2)'\beta),\\
\Gamma_1(z) &= \frac{d}{d\beta'}D_hQ(\beta,h_F)[h]\\
&= D_{\beta h}Q(\beta,h_F)[h]\\
&= \mathbb{E}_F[(X_1 - h_{2F}(X_2))'\beta(h_1(X_2) - h_2(X_2)'\beta]\\
&\quad + \mathbb{E}_F[(Y - h_{1F}(X_2) - (X_1 - h_{2F}(X_2))'\beta)h_2(X_2)]\\
&= 0,
\end{aligned}$$

where the last equality holds by the law of iterated expectations (i.e., first conditional on X_2). By Theorem 3.3 of Ichimura and Lee (2010), the influence function of an estimator $\hat{\beta}_n$ is

$$\begin{aligned}
\psi(z) &= -V_0^{-1}\Delta_1(z)\\
&= -\{\mathbb{E}[(X_1 - h_2(s(X_1,X_2))) \cdot (X_1 - h_2(s(X_1,X_2)))']\}^{-1}\\
&\quad \cdot [y - h_1(s(x_1,x_2)) - (x_1 - h_2(s(x_1,x_2)))'\beta] \cdot (x_1 - h_2(x_1,x_2)).
\end{aligned} \tag{D.1}$$

[4] Here $s(X_1,X_2) = s$ is a shorthand notation to indicate conditioning on all the additively separable components of $s(X_1,X_2)$. For example, in the model (4.2) in Section 4, $s(X_1,X_2) = s$ is a shorthand for the entire vector $(X_2', X_{11}X_{21}, X_{12}X_{22}, X_{13}X_{23}, X_{14}X_{24})'$ being fixed.

Note that $\Gamma_1(z)$, the term in Ichimura and Lee (2010) that captures the impact of first step estimation error of h on the asymptotic distribution of $\hat{\beta}_n$, is zero.

For the parametric estimator, s is restricted to be a linear function of x_2 only, i.e., $s(x_1, x_2) = \theta_0 + x_2'\theta_1$, then $\hat{\beta}_{n,P}$ is just the least squares coefficient of X_1 in a linear regression of Y on X_1, X_2, and an intercept. Under this modeling restriction, both h_1 and h_2 are also linear functions of x_2 only. Let $X_{2i}^* \equiv (1, X_{2i}')'$, then standard results of linear regressions imply that $h_{1F,P}(x_2) \equiv \mathbb{E}_F(Y|X_2 = x_2) = x_2^{*\prime}\gamma_{1F}$ and $h_{2F,P}(x_2) \equiv \mathbb{E}_F(X_1|X_2 = x_2) = x_2^{*\prime}\gamma_{2F}$ with $\gamma_{1F} = [\mathbb{E}_F(X_2^* X_2^{*\prime})]^{-1}\mathbb{E}_F(X_2^* Y)$ and $\gamma_{2F} = [\mathbb{E}_F(X_2^* X_2^{*\prime})]^{-1}\mathbb{E}_F(X_2^* X_1')$, which can then be plugged into (D.1) to obtain the influence function of $\hat{\beta}_{n,P}$. In order to get its sample version, let $\hat{\gamma}_{1,n} \equiv \left(\sum_{i=1}^n X_{2i}^* X_{2i}^{*\prime}\right)^{-1}\left(\sum_{i=1}^n X_{2i}^* Y_i\right)$ and $\hat{\gamma}_{2,n} \equiv \left(\sum_{i=1}^n X_{2i}^* X_{2i}^{*\prime}\right)^{-1}\left(\sum_{i=1}^n X_{2i}^* X_{1i}'\right)$, then one has

$$\begin{aligned}\psi_{n,P}(Z_i) = &-\left[\frac{1}{n}\sum_{i=1}^{n_1}(X_{1i} - X_{2i}^{*\prime}\hat{\gamma}_{2,n})(X_{1i} - X_{2i}^{*\prime}\hat{\gamma}_{2,n})'\right]^{-1} \\ &\cdot \{Y_i - X_{2i}^{*\prime}\hat{\gamma}_{1,n} - (X_{1i} - X_{2i}^{*\prime}\hat{\gamma}_{2,n})'\hat{\beta}_{n,SP}\} \cdot (X_{1i} - X_{2i}^{*\prime}\hat{\gamma}_{2,n}),\end{aligned} \tag{D.2}$$

where note that β in the influence function is replaced by its robust estimator $\hat{\beta}_{n,SP}$.

For the semiparametric estimator, a series of basis functions $G^L(x_1, x_2) \equiv (g_{1L}(x_1, x_2), g_{2L}(x_1, x_2), \ldots, g_{LL}(x_1, x_2))'$ is used to approximate the unknown function $s(x_1, x_2)$ in the original model, where L is an integer that increases with n and $g_{lL}(x_1, x_2)$ is a known function (e.g., polynomial functions) for each $l \in \{1, \ldots, L\}$. In this case, $\hat{\beta}_{n,SP}$ is just the least squares coefficient of X_1 in a linear regression of Y on X_1 and $G^L(X_1, X_2)$, and its influence function is what is in (D.1). The argument in Ackerberg, Chen and Hahn (2012) and Ackerberg, Chen, Hahn and Liao (2014) allows one to treat this series approximation as the true model in estimating the asymptotic variance of $\hat{\beta}_{n,SP}$. To proceed, let

$$\begin{aligned}\hat{\lambda}_{1,n} &\equiv \left(\sum_{i=1}^n G^L(X_{1i}, x_{2i})G^{L\prime}(X_{1i}, X_{2i})\right)^{-1}\left(\sum_{i=1}^n G^L(X_{1i}, x_{2i})Y_i\right), \text{ and} \\ \hat{\lambda}_{2,n} &\equiv \left(\sum_{i=1}^n G^L(X_{1i}, x_{2i})G^{L\prime}(X_{1i}, X_{2i})\right)^{-1}\left(\sum_{i=1}^n G^L(X_{1i}, x_{2i})X_{1i}'\right),\end{aligned}$$

then one has

$$\begin{aligned}\psi_{n,SP}(Z_i) = &-\left[\frac{1}{n}\sum_{i=1}^{n_1}(X_{1i} - G^{L\prime}(X_{1i}, X_{2i})\hat{\lambda}_{2,n})(X_{1i} - G^{L\prime}(X_{1i}, X_{2i})\hat{\lambda}_{2,n})'\right]^{-1} \\ &\cdot \{Y_i - G^{L\prime}(X_{1i}, X_{2i})\hat{\lambda}_{1,n} - (X_{1i} - G^{L\prime}(X_{1i}, X_{2i})\hat{\lambda}_{2,n})'\hat{\beta}_{n,SP}\} \\ &\cdot (X_{1i} - G^{L\prime}(X_{1i}, X_{2i})\hat{\lambda}_{2,n}).\end{aligned} \tag{D.3}$$

As a result, the averaging weight can be constructed by first plugging (D.2) and (D.3) into (C.1), (C.2) and (C.3), and then plugging the latter into (2.2).

Two points are worth emphasizing here. First, β naturally arises in the influence functions (D.1) and is invariant to how the conditional mean function h is modeled. Therefore, when computing the sample analogs of the influence functions using (D.2) and (D.3), β should be replaced by $\hat{\beta}_{n,SP}$, the estimator that is consistent regardless of whether the joint normality restriction is correctly specified or not. Second, the nuisance function h directly enters the influence function of $\hat{\beta}_n$. As a result, how h is modeled (by the linear function of x_2 only or without such restriction) affects the functional form of the influence functions, even though neither (D.2) nor (D.3) contains a correction term for the first step estimation error of h.

Primitive conditions. For a specific model and specific estimators $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$, Conditions 1 and 3 are straightforward to verify, and Condition 2 can be verified under more primitive conditions. The following condition is the primitive condition of Condition 2 for the partially linear model 4.

Condition 2′. *Let $\|\cdot\|$ indicate the Euclidean norm of a vector and let $\|\cdot\|_{\mathcal{L}_2}$ indicate the $\mathcal{L}_2$ norm of a function. For the partially linear model in* (4.1) *and the estimators described in Section 2, assume that the following conditions hold for any $F \in \mathcal{F}$, where $0 < M < \infty$ and $\tau > 0$ are some generic constants.*

(i) $\mathbb{E}_F\{[X_1 - \mathbb{E}_F(X_1|s(X_1,X_2))] \cdot [X_1 - \mathbb{E}_F(X_1|s(X_1,X_2))]'\}$ *is positive definite.*

(ii) $\mathbb{E}_F\{[U + s(X_1,X_2) - \mathbb{E}_F(s(X_1,X_2)|X_2)]^2[X_1 - \mathbb{E}_F(X_1|X_2)] \cdot [X_1 - \mathbb{E}_F(X_1|X_2)]'\}$ *is positive definite.*

(iii) For functions $s(x_1,x_2)$ and $h_{2F,P}(x_2) \equiv \mathbb{E}_F(X_1|X_2) = x_2^{*\prime}\gamma_{2F}$, there exist d_s, d_2, $\pi_{s,L}$ and $\pi_{2,L}$ such that $\|s(x_1,x_2) - G^{L\prime}(x_1,x_2)\pi_{s,L}\|_{\mathcal{L}_2} = O(L^{-d_s})$ and $\|h_{2F}(x_2) - G^{L\prime}(x_2)\pi_{2,L}\|_{\mathcal{L}_2} = O(L^{-d_2})$ as $L \to \infty$, where $G^L(x_1,x_2)$ and $G^L(x_2)$ are series basis functions of order L.

(iv) $var_F[Y|X_1, s(X_1,X_2)] \le M < \infty$ *and* $var_F[X_1|s(X_1,X_2)] \le M$.

(v) $\mathbb{E}_F\{\|X_1 - \mathbb{E}_F[X_1|s(X_1,X_2)]\|^{2+\tau}\} \le M$.

(vi) The series order L is such that $L \to \infty$, $L/n \to 0$ and $\sqrt{n}L^{-d_s-d_2} \to 0$ as $n \to 0$.

(vii) The variance-covariance matrix (under F) of $(X_1', X_2')'$ is positive definite.

(viii) $\mathbb{E}_F[\|X_1\|^{2+\tau}] \le M$ *and* $\mathbb{E}_F[\|X_2\|^{2+\tau}] \le M$.

Condition 2′(i) is the key identification requirement of β_F. Condition 2′(i) - (vi) follow Assumption 2 and Theorem 2 in Donald and Newey (1994), which ensure the asymptotic normality of $\hat{\beta}_{n,SP}$ based on series first step. Among them, (i) and (ii) ensure that the asymptotic variance-covariance matrix of $\hat{\beta}_{n,SP}$ is well defined. (iii) implies that the nuisance functions can be approximated well by the series basis functions; (iv) implies that they can be consistently estimated; (v) is the moment condition required by the central limit theorem; and (vi) gives the under-smoothing order of the series basis functions. (vii) is the key identification requirement of $\beta_{F,P}$, and (viii) is the usual moment condition for the asymptotic normality of $\hat{\beta}_{n,P}$ based on linear regression of Y on $(X_1', X_2')'$.[5]

Verification of the primitive conditions. This part verifies Conditions 1, 2′ and 3 for the Monte Carlo model (4.2) and the estimators used in Section 4. Recall that in this model,

$$s(x_1,x_2) \equiv x_2'\theta_1 + t(x_1,x_2), \text{ with } t(x_1,x_2) \equiv \rho\left(\sum_{j=1}^{4}\theta_{2j}\exp(x_{2j}) + \sum_{j=1}^{4}\theta_{3j}x_{1j}x_{2j}\right). \tag{D.4}$$

Because the misspecified model (4.3) only uses $\theta_0 + x_2'\theta_1$ as $s(x_1,x_2)$, ρ controls the degree of misspecification. Note that for $\forall F \in \mathcal{F}$, the probability limit of the parametric least squares estimator $(\theta_{0,F,P}', \beta_{F,P}', \theta_{1,F,P}')'$ satisfies

$$\begin{bmatrix} \theta_{0,F,P} - \theta_{0,F} \\ \beta_{F,P} - \beta_F \\ \theta_{1,F,P} - \theta_{1,F} \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}_F(X_1') & \mathbb{E}_F(X_2') \\ \mathbb{E}_F(X_1) & \mathbb{E}_F(X_1X_1') & \mathbb{E}_F(X_1X_2') \\ \mathbb{E}_F(X_2) & \mathbb{E}_F(X_2X_1') & \mathbb{E}_F(X_2X_2') \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}_F[t(X_1,X_2)] \\ \mathbb{E}_F[X_1t(X_1,X_2)] \\ \mathbb{E}_F[X_2t(X_1,X_2)] \end{bmatrix}, \tag{D.5}$$

where $(\theta_{0,F,P}', \beta_{F,P}', \theta_{1,F,P}')'$ is the pseudo-true parameter value in the misspecified model (4.3) and $(\theta_{0,F}', \beta_F', \theta_{1,F}')'$ is the true parameter vector in model (4.2).[6] The joint normal distribution of $(X_1', X_2')'$ implies that on the right hand side of (D.5), the entries of the first matrix are non-zero finite numbers, and the

[5] The joint asymptotic normality can be shown under Condition 2′ by invoking Cramér-Wold theorem (not provided here).

[6] In particular, $\theta_{0,F} = 0$.

second vector is proportional to ρ, since $\mathbb{E}_F(X_l X_{1j} X_{2j})$ and $\mathbb{E}_F[X_l \exp(X_{2j})]$ are both finite ($l = 1, 2$ and $j = 1, 2, 3, 4$).[7] So, one gets

$$\delta_F \equiv \beta_{F,P} - \beta_F = c_1 \rho \tag{D.6}$$

with non-zero c_1, where the non-zero constant c_1 depends on the moments of polynomials up to the third order and the exponential functions of $(X_1', X_2')'$. As a result, as long as the values of ρ contain an open set around 0, Condition 1(ii) is satisfied. Moreover, note that the nuisance function h_F in Condition 1(i) is $s(x_1, x_2)$ and g_{γ_F} in Condition 1(i) is $x_2' \theta_{1,F,P}$, then one has

$$\|g_{\gamma_F} - h_F\|_{\mathcal{L}_2} = \|x_2'(\theta_{1,F,P} - \theta_{1,F}) + \theta_{0,F,P} - t(x_1, x_2)\|_{\mathcal{L}_2} = c_2|\rho|, \text{ for some } c_2 > 0,$$

where the second equality holds due to (D.5), the definition of $t(x_1, x_2)$ in (D.4), the joint normal distribution of $(X_1', X_2')'$ and that $\theta_{0,F} = 0$ in model (4.2). As a result, Condition 1(i) is satisfied.

The joint normal distribution of $(X_1', X_2')'$ in Section 4 immediately implies that Conditions 2′(vii), (viii) and the second part of (iv) are satisfied.[8] Moreover, the normal distribution of U and its independence from $(X_1', X_2')'$ ensure Condition 2′ (ii) and the first part of (iv). In addition, the definition of $s(x_1, x_2)$ in (D.4) shows that X_1 and $s(X_1, X_2)$ are not perfectly collinear, indicating that Condition 2′(i) and (v) are satisfied. Furthermore, note that $s(x_1, x_2)$ only contains linear, quadratic and exponential functions of x_1 and x_2 and $h_{2F,P}(x_2)$ only contains linear function of x_2, which are all four times continuously differentiable functions, so Condition 2′(iii) is satisfied with $d_s = d_2 = \frac{1}{2}$.[9] Given this, one can choose L such that $L \to \infty$, $L/n \to 0$ and $L^2/n \to \infty$ to satisfy Condition 2′(vi).

To verify Condition 3, first recall $\delta_F = c_1 \rho$ with non-zero c_1 shown in (D.6), so δ_F belongs to a compact set as long as ρ does. Second, note that the asymptotic variance-covariance matrices of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are

$$V_{F,SP} = \sigma_U^2 \cdot \left(\mathbb{E}_F\{[X_1 - \mathbb{E}_F(X_1|s(X_1, X_2))] \cdot [X_1 - \mathbb{E}_F(X_1|s(X_1, X_2))]'\}\right)^{-1}, \tag{D.7}$$

$$\begin{aligned} V_{F,P} = {} & \left(\mathbb{E}_F\{[X_1 - \mathbb{E}_F(X_1|X_2)] \cdot [X_1 - \mathbb{E}_F(X_1|X_2)]'\}\right)^{-1} \\ & \cdot \left(\mathbb{E}_F\{[U + s(X_1, X_2) - \mathbb{E}_F(s(X_1, X_2)|X_2)]^2 [X_1 - \mathbb{E}_F(X_1|X_2)] \cdot [X_1 - \mathbb{E}_F(X_1|X_2)]'\}\right) \\ & \cdot \left(\mathbb{E}_F\{[X_1 - \mathbb{E}_F(X_1|X_2)] \cdot [X_1 - \mathbb{E}_F(X_1|X_2)]'\}\right)^{-1}, \end{aligned} \tag{D.8}$$

$$\begin{aligned} C_F = {} & \left(\mathbb{E}_F\{[X_1 - \mathbb{E}_F(X_1|s(X_1, X_2))] \cdot [X_1 - \mathbb{E}_F(X_1|s(X_1, X_2))]'\}\right)^{-1} \\ & \cdot \left(\mathbb{E}_F\{[U + s(X_1, X_2) - \mathbb{E}_F(s(X_1, X_2)|X_2)] \cdot [X_1 - \mathbb{E}_F(X_1|s(X_1, X_2))]'\}\right) \\ & \cdot \left(\mathbb{E}_F\{[X_1 - \mathbb{E}_F(X_1|X_2)] \cdot [X_1 - \mathbb{E}_F(X_1|X_2)]'\}\right)^{-1}, \end{aligned} \tag{D.9}$$

where $\sigma_U^2 \equiv \mathbb{E}_F(U^2)$. Given the specification in (4.2) (reiterated in (D.4)) and the joint normal distribution of $(X_1', X_2')'$ in Section 4, the following points can be verified.

1. $h_{2F,P}(\cdot) = \mathbb{E}_F(X_1|X_2)$ defined before (D.2) is a 4×1 vector-valued linear function of x_2 that does not depend on ρ; in particular, its jth coordinate is $h_{2F,P,j}(x_2) \equiv \mathbb{E}_F(X_{1j}|X_2 = x_2) = 0.4 + 0.2 \sum_{l=1}^4 x_{2,l}$ for $j = 1, 2, 3, 4$.

2. Note that once the values of $\exp(X_{2j})$ and $X_{1j} X_{2j}$ ($j = 1, 2, 3, 4$) are fixed, then so are the values of X_{1j} and X_{2j} ($j = 1, 2, 3, 4$); and vice versa. For this reason, the function $h_{2F}(\cdot) = \mathbb{E}_F(X_1|s(X_1, X_2)) = \mathbb{E}_F(X_1|X_2, X_{11}X_{21}, X_{12}X_{22}, X_{13}X_{23}, X_{14}X_{24})$ defined before (D.1) does not depend on ρ, although its

[7] The finite moments of $\exp(X_2)$ can be shown using the moment generating function of the normally distributed X_2. For example, $\mathbb{E}\{[\exp(X_{2j})]^2\} = \mathbb{E}[\exp(2X_{2j})] = M_{X_{2j}}(2) = \exp(2\mu_{2j} + 2\sigma_{2j}^2) < \infty$, with $\mu_{2j} = 2$ and $\sigma_{2j}^2 = 0.5^2$.

[8] Note that conditional variance is bounded above by unconditional variance.

[9] $s(x_1, x_2)$ is four times continuously differentiable and has eight arguments, so by the discussion that follows Assumption 3 in Newey (1997), $d_s = \frac{4}{8} = \frac{1}{2}$. Similarly, $h_{2F}(x_2)$ is twice continuously differentiable and has four arguments, so $d_2 = \frac{2}{4} = \frac{1}{2}$.

functional form is difficult to obtain and hence is omitted here.[10]

3. By the specification in (4.2), one has

$$s(X_1, X_2) - \mathbb{E}_F(s(X_1, X_2)|X_2) = \rho \sum_{j=1}^{4} \theta_{3j} X_{2j}[X_{1j} - \mathbb{E}_F(X_{1j}|X_2)] \equiv C_3\rho, \tag{D.10}$$

where C_3 is a random variable that depends on X_1 and X_2.

Point 2 immediately implies that $V_{F,SP}$ in (D.7) equals to

$$\begin{aligned} V_{F,SP} &= \sigma_U^2 \cdot [\mathbb{E}_F(W_{F,SP}W_{F,SP}')]^{-1}, \\ &\text{with } W_{F,SP} \equiv X_1 - \mathbb{E}_F(X_1|X_2, X_{11}X_{21}, X_{12}X_{22}, X_{13}X_{23}, X_{14}X_{24}), \end{aligned} \tag{D.11}$$

which does not depend on ρ. Points 1 - 3 together imply that $V_{F,P}$ in (D.8) equals to

$$\begin{aligned} V_{F,P} &= [\mathbb{E}_F(W_{F,P}W_{F,P}')]^{-1} \cdot \{\mathbb{E}_F[(U + C_6\rho)^2 W_{F,P}W_{F,P}']\} \cdot [\mathbb{E}_F(W_{F,P}W_{F,P}')]^{-1}, \\ &\text{with } W_{F,P} \equiv X_1 - \mathbb{E}_F(X_1|X_2), \end{aligned} \tag{D.12}$$

which is a quadratic function of ρ. Similarly, one can show that C_F in (D.9) is a linear function of ρ. In summary, $\bar{S}(F)$ defined in (3.17) is a quadratic function of ρ; that is, for $F \in \mathcal{F}$ such that $\delta_F = c_1\rho$, there exist some fixed vectors c_4, c_5 and c_6 such that $\bar{S}(F) = c_4 + c_5\rho + c_6\rho^2$. This has two implications. First, as long as ρ takes values from a compact set, so does $S(F)$ defined in (3.17), satisfying Condition 3(i). Second, Condition 3(ii) is satisfied with $\kappa = 2$, $\epsilon_F = 1$ and C being some constant depending on c_4, c_5 and c_6 but not ρ.

Verification of $V_{F,SP} \geq V_{F,P}$ in Condition 2(i). The paragraph that follows Condition 2 provided intuition for $V_{F,SP} \geq V_{F,P}$ using the semiparametric efficiency bound and Le Cam's third lemma. For a specific model, however, this result can often be verified directly. What follows will use (D.11) and (D.12) to verify it for the parameterization of the partially linear model in (4.2).

Recall that (D.6) shows that $\delta_F \equiv \beta_{F,P} - \beta_F = c_1\rho$ with non-zero c_1. Also recall that the presumption of Condition 2(i) is $\|d\| < \infty$ with $n^{1/2}\delta_{F_n} \to d$, then any sequence ρ_n of ρ values considered here satisfies $\rho_n = \frac{\delta_{F_n}}{c_1} = O(n^{-1/2})$. Moreover, U is independent of $W_{F,P}$ due to the independence between U and $(X_1', X_2')'$. These together imply that in the scenario of Condition 2(i), $V_{F,P}$ in (D.12) equals to $\sigma_U^2 \cdot [\mathbb{E}_F(W_{F,P}W_{F,P}')]^{-1}$. Comparing it with (D.11), it is obvious that $\mathbb{E}_F(W_{F,SP}W_{F,SP}') \leq \mathbb{E}_F(W_{F,P}W_{F,P}')$ since $W_{F,SP}$ conditions on more variables. This further implies that $V_{F,SP} \geq V_{F,P}$.

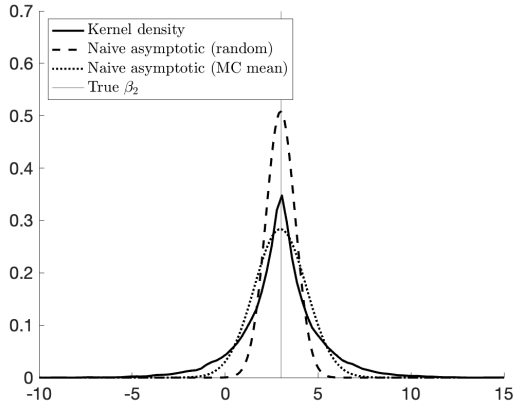
More Monte Carlo results. This subsection reports Monte Carlo results for the second, the third and the fourth coordinates of β (i.e., β_2, β_3 and β_4). Similar results for β_1, the first coordinate of β, are reported in Figure 2 and Table 1 in Section 4.

Figures D.1 - D.3 plot the Monte Carlo distributions (kernel densities) of the averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ (thick solid lines) for representative ρ values. Figure D.1 is for β_2, Figure D.2 is for β_3 and Figure D.3 is for β_4. In the same figures, the normal distributions based on the naive inference method with the common standard error are represented by the thick dashed lines (one randomly chosen Monte Carlo replication) and dotted lines (averaged over all Monte Carlo replications). It is obvious that the naive inference method

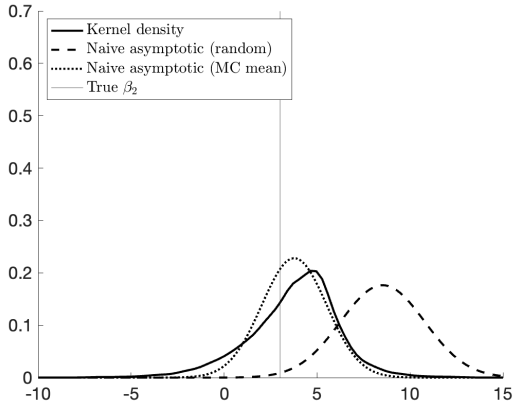
[10] See Footnote 4 for details on $\mathbb{E}_F(X_1|s(X_1, X_2))$.

Figure D.1: True vs. Naive Distributions of $\hat{\beta}_{n,\hat{w}_n,2}$ for the Partially Linear Model

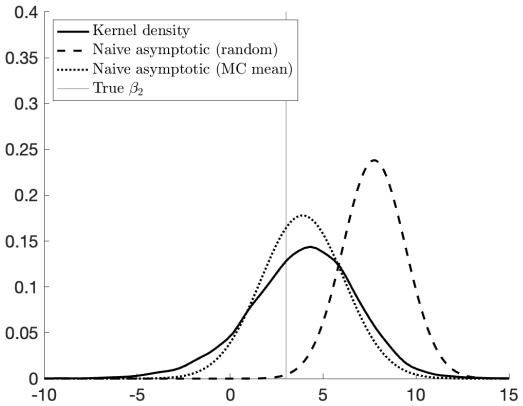
(a) Correct Specification ($\rho = 0$):



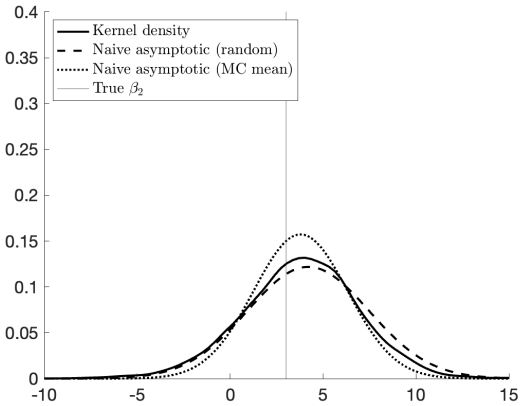
(b) Misspecification ($\rho = 0.2$):



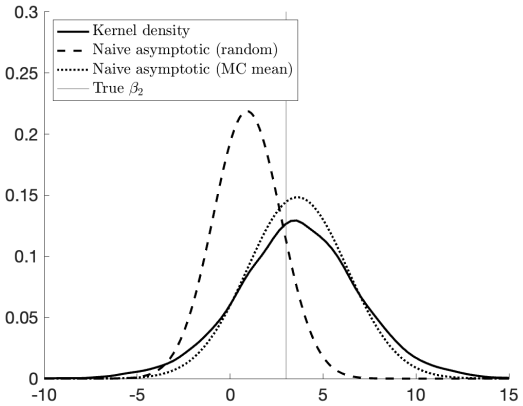
(c) Misspecification ($\rho = 0.4$):



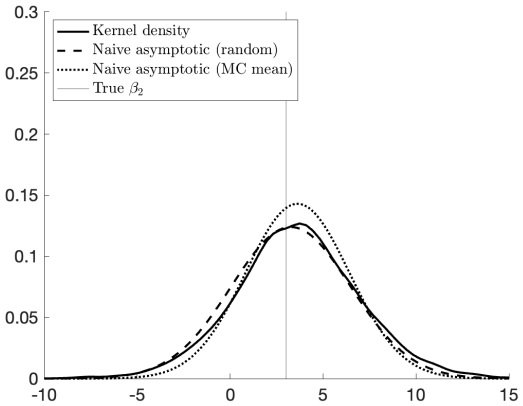
(d) Misspecification ($\rho = 0.6$):



(e) Misspecification ($\rho = 0.8$):



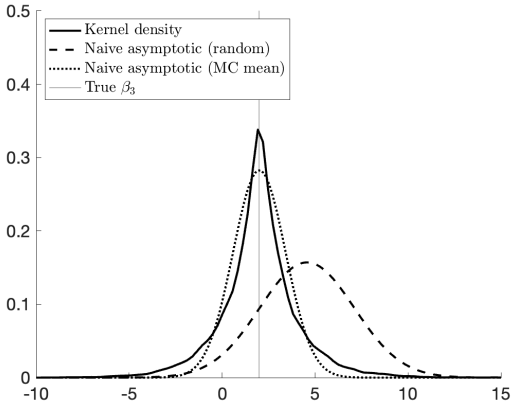
(f) Misspecification ($\rho = 1$):



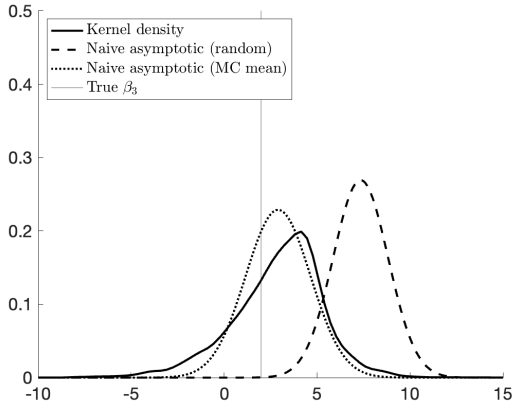
Notes: (1) All distributions are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size.
(2) Solid lines represent the MC distributions of $\hat{\beta}_{n,\hat{w}_n,2}$, the averaging estimator of β_2. Dashed and dotted lines both represent the asymptotic distribution of $\hat{\beta}_{n,\hat{w}_n,2}$ if the naive inference method, which takes $\hat{w}_n$ as fixed, is used. The former show a randomly chosen MC replication, while the latter show the average over all MC replications.
(3) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Figure D.2: True vs. Naive Distributions of $\hat{\beta}_{n,\hat{w}_n,3}$ for the Partially Linear Model

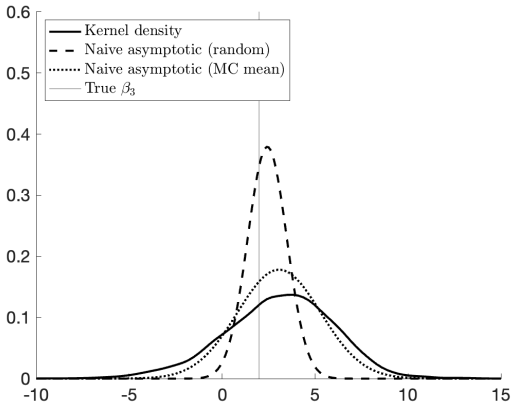
(a) Correct Specification ($\rho = 0$):



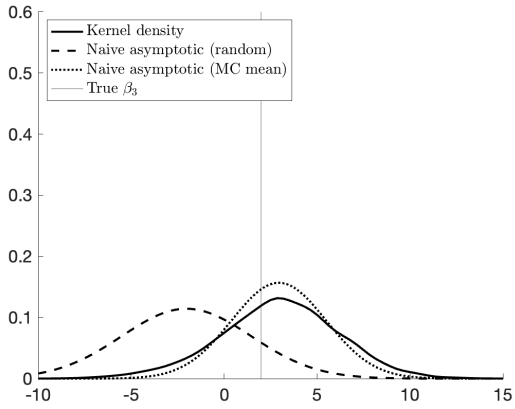
(b) Misspecification ($\rho = 0.2$):



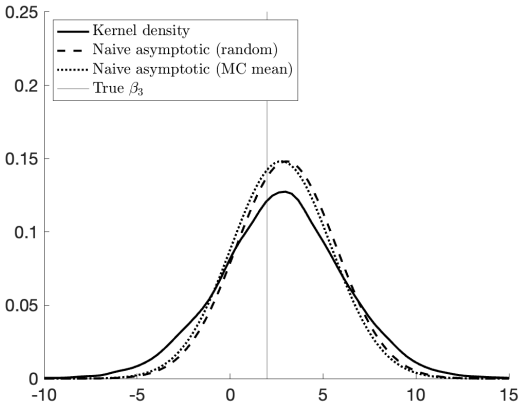
(c) Misspecification ($\rho = 0.4$):



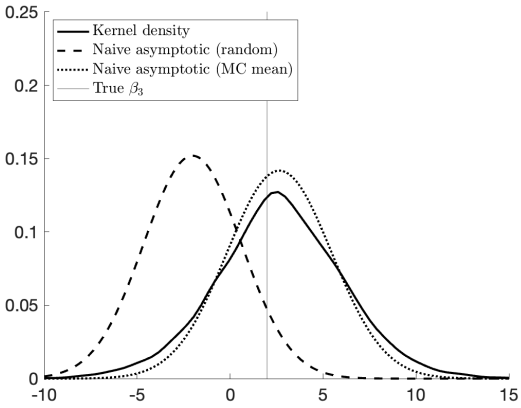
(d) Misspecification ($\rho = 0.6$):



(e) Misspecification ($\rho = 0.8$):



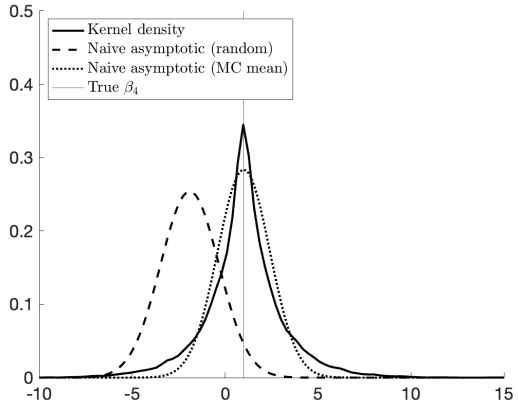
(f) Misspecification ($\rho = 1$):



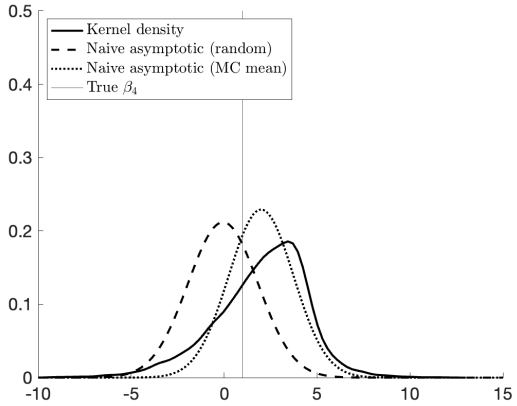
Notes: (1) All distributions are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size.
(2) Solid lines represent the MC distributions of $\hat{\beta}_{n,\hat{w}_n,3}$, the averaging estimator of β_3. Dashed and dotted lines both represent the asymptotic distribution of $\hat{\beta}_{n,\hat{w}_n,3}$ if the naive inference method, which takes $\hat{w}_n$ as fixed, is used. The former show a randomly chosen MC replication, while the latter show the average over all MC replications.
(3) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Figure D.3: True vs. Naive Distributions of $\hat{\beta}_{n,\hat{w}_n,4}$ for the Partially Linear Model

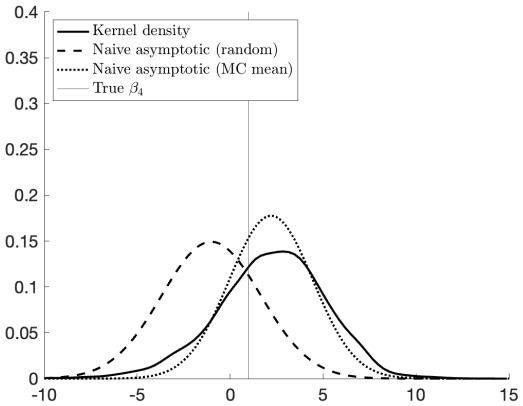
(a) Correct Specification ($\rho = 0$):



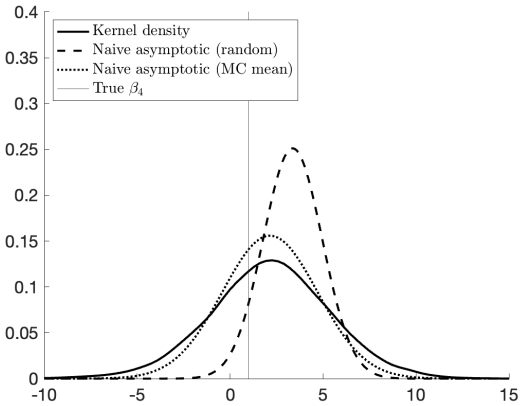
(b) Misspecification ($\rho = 0.2$):



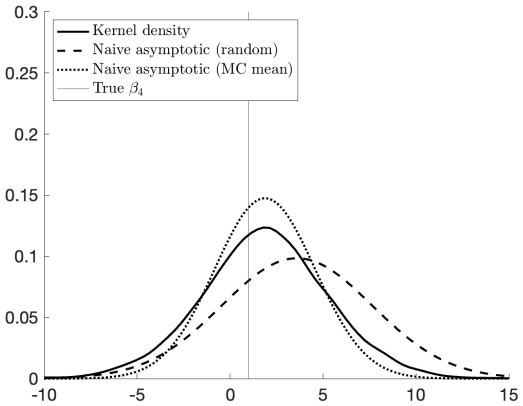
(c) Misspecification ($\rho = 0.4$):



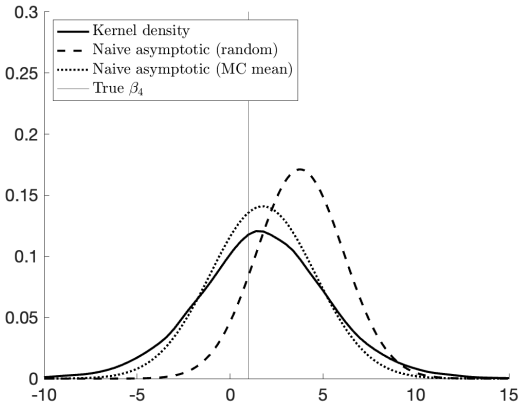
(d) Misspecification ($\rho = 0.6$):



(e) Misspecification ($\rho = 0.8$):



(f) Misspecification ($\rho = 1$):



Notes: (1) All distributions are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size.
(2) Solid lines represent the MC distributions of $\hat{\beta}_{n,\hat{w}_n,4}$, the averaging estimator of β_4. Dashed and dotted lines both represent the asymptotic distribution of $\hat{\beta}_{n,\hat{w}_n,4}$ if the naive inference method, which takes $\hat{w}_n$ as fixed, is used. The former show a randomly chosen MC replication, while the latter show the average over all MC replications.
(3) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Table D.1: Rejection Rates for $\hat{\beta}_{n,\hat{w}_n,2}$ in the Partially Linear Model (5% Level)

ρ	$\hat{\beta}_{n,SP,2}$		$\hat{\beta}_{n,\hat{w}_n,2}$						CI Length
			Naive		Naive (robust SE)		Two-step		
	Size	Power	Size	Power	Size	Power	Size	Power	$\frac{CI(\hat{\beta}_{n,\hat{w}_n,2})}{CI(\hat{\beta}_{n,SP,2})}$
0.00	9.19%	24.71%	9.20%	64.11%	9.10%	63.93%	1.55%	10.96%	33.1069
0.05	10.03%	25.49%	13.89%	66.46%	13.60%	66.38%	1.73%	14.32%	33.1784
0.10	9.36%	23.99%	18.18%	65.45%	17.99%	65.38%	1.77%	16.07%	33.2881
0.15	8.94%	24.48%	21.56%	63.96%	21.39%	63.89%	2.01%	18.84%	33.4046
0.20	9.80%	24.67%	22.22%	61.40%	22.06%	61.36%	2.48%	22.26%	33.5380
0.25	9.34%	24.71%	21.00%	58.04%	20.96%	57.98%	2.86%	23.80%	33.6886
0.30	9.93%	24.78%	20.19%	53.85%	20.10%	53.83%	3.67%	25.05%	33.8610
0.35	9.56%	24.86%	19.26%	51.33%	19.21%	51.24%	4.08%	26.55%	33.9920
0.40	9.85%	25.03%	17.26%	48.73%	17.21%	48.71%	4.71%	26.65%	34.1516
0.45	9.11%	23.67%	15.68%	44.38%	15.67%	44.35%	4.92%	25.28%	34.2996
0.50	9.99%	24.93%	15.83%	43.27%	15.79%	43.22%	5.67%	26.60%	34.4368
0.55	9.95%	24.79%	14.70%	41.33%	14.65%	41.29%	5.54%	26.01%	34.5620
0.60	9.43%	24.07%	13.32%	39.86%	13.32%	39.80%	5.57%	25.21%	34.6631
0.65	9.64%	24.30%	12.89%	37.71%	12.88%	37.73%	6.00%	24.58%	34.7839
0.70	9.74%	24.48%	13.03%	36.81%	13.04%	36.81%	6.25%	24.56%	34.8843
0.75	9.68%	24.40%	12.19%	35.12%	12.15%	35.11%	6.10%	23.76%	34.9779
0.80	10.40%	24.34%	12.28%	34.57%	12.28%	34.56%	6.40%	23.90%	35.0621
0.85	9.94%	24.49%	11.89%	34.28%	11.89%	34.28%	6.40%	23.52%	35.1280
0.90	9.93%	24.92%	11.90%	33.84%	11.88%	33.82%	6.41%	23.66%	35.2114
0.95	9.93%	24.16%	11.43%	32.50%	11.42%	32.52%	6.31%	22.65%	35.2659
1.00	9.69%	24.71%	11.30%	32.19%	11.30%	32.18%	6.24%	22.80%	35.3216
1.05	9.98%	25.03%	11.41%	32.58%	11.41%	32.55%	6.43%	22.92%	35.3794
1.10	10.46%	24.29%	11.98%	31.07%	11.98%	31.05%	6.75%	21.94%	35.4232
1.15	9.80%	24.67%	10.98%	31.84%	10.95%	31.85%	6.59%	22.27%	35.4762
1.20	10.11%	24.46%	10.77%	30.80%	10.77%	30.79%	6.23%	21.67%	35.5112
1.25	10.22%	23.85%	11.30%	29.97%	11.30%	29.96%	6.68%	21.26%	35.5386
1.30	10.43%	24.41%	11.34%	30.03%	11.33%	30.02%	6.41%	21.53%	35.5749

Notes: (1) This table reports the inference results for $\hat{\beta}_{n,\hat{w}_n,2}$, the averaging estimator of β_2.
(2) All results are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size. The two-step inference method uses $S = 1000$ replications to simulate the distribution of $\bar{\xi}_{F,d} \equiv (1 - w_F)\xi_{F,SP} + w_F(\xi_{F,P} + d)$ in (3.22).
(3) The naive inference methods treat the averaging weight $\hat{w}_n$ as non-random, and hence underestimate the randomness in $\hat{\beta}_{n,\hat{w}_n,2}$. Two naive methods are reported here: the first uses the common estimators of $V_{F,P}$ and C_F, which might be biased under misspecification (see the discussion after (C.3) for details); and the second combines $\widehat{V}_{n,P}$ and $\widehat{C}_n$ given in (C.2) and (C.3) with the robust influence function $\psi_{n,P}(z)$ in Appendix D, which are robust under misspecification (robust SE).
(4) The test value for the "Size" columns is 3, the true value of β_2; the test value for the "Power" columns is 0.
(5) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Table D.2: Rejection Rates for $\hat{\beta}_{n,\hat{w}_n,3}$ in the Partially Linear Model (5% Level)

ρ	$\hat{\beta}_{n,SP,3}$		$\hat{\beta}_{n,\hat{w}_n,3}$						CI Length
			Naive		Naive (robust SE)		Two-step		
	Size	Power	Size	Power	Size	Power	Size	Power	$\frac{CI(\hat{\beta}_{n,\hat{w}_n,3})}{CI(\hat{\beta}_{n,SP,3})}$
0.00	9.35%	16.49%	9.42%	46.84%	9.22%	46.65%	1.80%	6.06%	31.8093
0.05	9.69%	16.52%	14.45%	52.31%	14.17%	52.17%	1.93%	8.16%	31.8198
0.10	9.49%	16.33%	20.92%	53.27%	20.61%	53.14%	2.20%	10.44%	31.8614
0.15	10.07%	16.77%	25.42%	52.54%	25.21%	52.52%	2.78%	13.99%	31.9477
0.20	9.63%	17.04%	25.25%	50.24%	25.09%	50.21%	3.20%	17.46%	31.0517
0.25	9.65%	17.02%	23.98%	47.67%	23.86%	47.65%	4.42%	20.00%	32.1615
0.30	10.15%	16.74%	22.78%	43.62%	22.56%	43.59%	5.70%	20.86%	32.3306
0.35	9.61%	16.22%	20.77%	40.84%	20.76%	40.77%	6.25%	21.78%	32.4561
0.40	9.90%	16.80%	19.42%	38.84%	19.35%	38.83%	7.05%	23.04%	32.6202
0.45	10.06%	16.79%	18.22%	35.95%	18.20%	35.87%	7.97%	22.56%	32.7834
0.50	10.25%	15.83%	16.34%	32.80%	16.28%	32.77%	7.71%	21.51%	32.9379
0.55	9.89%	16.35%	15.58%	31.71%	15.56%	31.67%	8.12%	21.49%	33.0857
0.60	10.37%	16.99%	15.40%	30.73%	15.35%	30.74%	8.68%	21.88%	33.2150
0.65	9.91%	16.83%	14.31%	29.00%	14.30%	29.00%	8.37%	20.75%	33.3506
0.70	10.14%	17.00%	13.90%	28.56%	13.86%	28.51%	8.43%	20.84%	33.4695
0.75	9.46%	16.54%	12.88%	26.52%	12.87%	26.50%	8.05%	19.66%	33.5899
0.80	10.62%	17.09%	13.44%	26.64%	13.42%	26.60%	8.71%	20.37%	33.6846
0.85	9.89%	16.79%	12.35%	25.31%	12.35%	25.31%	8.05%	19.32%	33.7756
0.90	10.71%	17.31%	12.88%	25.08%	12.87%	25.10%	8.72%	19.34%	33.8567
0.95	10.64%	16.45%	12.50%	23.82%	12.49%	23.81%	8.11%	18.27%	33.9313
1.00	10.08%	16.81%	12.13%	23.45%	12.12%	23.44%	8.02%	18.17%	33.9944
1.05	10.93%	16.74%	12.61%	23.32%	12.61%	23.30%	8.63%	18.13%	34.0518
1.10	10.58%	17.66%	12.37%	23.85%	12.36%	23.84%	8.58%	18.72%	34.1105
1.15	10.69%	17.19%	12.31%	22.54%	12.32%	22.53%	8.35%	18.01%	34.1666
1.20	10.91%	17.07%	12.18%	22.46%	12.17%	22.41%	8.71%	17.61%	34.2091
1.25	11.08%	17.57%	12.51%	22.91%	12.52%	22.93%	8.63%	17.98%	34.2507
1.30	11.08%	17.38%	12.50%	21.84%	12.49%	21.86%	8.75%	17.59%	34.2906

Notes: (1) This table reports the inference results for $\hat{\beta}_{n,\hat{w}_n,3}$, the averaging estimator of β_3.
(2) All results are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size. The two-step inference method uses $S = 1000$ replications to simulate the distribution of $\bar{\xi}_{F,d} \equiv (1 - w_F)\xi_{F,SP} + w_F(\xi_{F,P} + d)$ in (3.22).
(3) The naive inference methods treat the averaging weight $\hat{w}_n$ as non-random, and hence underestimate the randomness in $\hat{\beta}_{n,\hat{w}_n,3}$. Two naive methods are reported here: the first uses the common estimators of $V_{F,P}$ and C_F, which might be biased under misspecification (see the discussion after (C.3) for details); and the second combines $\widehat{V}_{n,P}$ and $\widehat{C}_n$ given in (C.2) and (C.3) with the robust influence function $\psi_{n,P}(z)$ in Appendix D, which are robust under misspecification (robust SE).
(4) The test value for the "Size" columns is 2, the true value of β_3; the test value for the "Power" columns is 0.
(5) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

Table D.3: Rejection Rates for $\hat{\beta}_{n,\hat{w}_n,4}$ in the Partially Linear Model (5% Level)

ρ	$\hat{\beta}_{n,SP,4}$		$\hat{\beta}_{n,\hat{w}_n,4}$						CI Length
			Naive		Naive (robust SE)		Two-step		
	Size	Power	Size	Power	Size	Power	Size	Power	$\frac{CI(\hat{\beta}_{n,\hat{w}_n,4})}{CI(\hat{\beta}_{n,SP,4})}$
0.00	10.25%	11.89%	10.34%	24.59%	10.19%	24.24%	1.61%	2.57%	33.1842
0.05	9.75%	11.86%	15.86%	34.95%	15.67%	34.70%	2.00%	3.43%	33.1979
0.10	9.60%	11.69%	23.23%	38.97%	23.00%	38.78%	1.97%	4.47%	33.2342
0.15	9.61%	11.45%	27.28%	40.30%	27.13%	40.29%	2.58%	5.88%	33.2914
0.20	9.26%	11.05%	27.23%	39.53%	27.16%	39.41%	3.19%	8.01%	33.3738
0.25	9.52%	11.28%	26.18%	36.56%	26.08%	36.48%	4.38%	10.66%	33.4788
0.30	10.48%	11.81%	25.09%	34.94%	25.09%	34.85%	6.25%	13.17%	33.6082
0.35	10.01%	11.59%	22.91%	32.41%	22.87%	32.35%	7.43%	14.00%	33.7167
0.40	8.75%	10.72%	19.98%	29.64%	19.96%	28.59%	7.45%	14.08%	33.8650
0.45	9.77%	11.76%	19.52%	26.84%	19.54%	26.77%	8.63%	15.06%	34.0064
0.50	9.53%	11.68%	18.34%	25.77%	18.33%	25.71%	8.81%	15.00%	34.1383
0.55	10.01%	11.73%	17.10%	23.62%	17.09%	23.57%	8.96%	14.76%	34.2743
0.60	9.84%	11.79%	16.38%	23.08%	16.38%	23.08%	8.82%	14.94%	34.3891
0.65	10.58%	12.45%	15.89%	21.82%	15.86%	21.77%	9.62%	14.52%	34.5273
0.70	9.89%	11.50%	14.72%	20.47%	14.67%	20.49%	8.77%	13.84%	34.6379
0.75	10.26%	12.13%	14.46%	20.10%	14.45%	20.11%	8.69%	13.84%	34.7625
0.80	10.48%	12.44%	14.12%	19.43%	14.11%	19.44%	9.18%	13.83%	34.8632
0.85	10.07%	11.54%	13.32%	18.49%	13.30%	18.47%	8.36%	12.68%	34.9520
0.90	11.21%	12.84%	14.21%	18.65%	14.20%	18.65%	9.07%	13.43%	35.0439
0.95	11.84%	13.21%	14.45%	18.67%	14.43%	18.66%	9.36%	13.47%	35.1175
1.00	10.83%	12.16%	13.06%	17.49%	13.03%	17.49%	8.21%	12.65%	35.1874
1.05	11.01%	12.42%	13.15%	17.24%	13.15%	17.25%	8.79%	12.42%	35.2540
1.10	11.90%	13.29%	13.68%	17.32%	13.66%	17.32%	8.80%	12.65%	35.3121
1.15	11.26%	12.32%	12.76%	16.22%	12.77%	16.22%	8.39%	11.92%	35.3659
1.20	11.22%	12.93%	12.98%	17.44%	12.99%	17.43%	8.58%	12.29%	35.4167
1.25	11.12%	12.97%	12.85%	16.68%	12.84%	16.67%	8.78%	12.54%	35.4533
1.30	11.85%	13.16%	12.97%	16.79%	13.00%	16.79%	8.77%	12.14%	35.4980

Notes: (1) This table reports the inference results for $\hat{\beta}_{n,\hat{w}_n,4}$, the averaging estimator of β_4.
(2) All results are based on $R = 10000$ Monte Carlo replications and $n = 1000$ sample size. The two-step inference method uses $S = 1000$ replications to simulate the distribution of $\bar{\xi}_{F,d} \equiv (1 - w_F)\xi_{F,SP} + w_F(\xi_{F,P} + d)$ in (3.22).
(3) The naive inference methods treat the averaging weight $\hat{w}_n$ as non-random, and hence underestimate the randomness in $\hat{\beta}_{n,\hat{w}_n,4}$. Two naive methods are reported here: the first uses the common estimators of $V_{F,P}$ and C_F, which might be biased under misspecification (see the discussion after (C.3) for details); and the second combines $\widehat{V}_{n,P}$ and $\widehat{C}_n$ given in (C.2) and (C.3) with the robust influence function $\psi_{n,P}(z)$ in Appendix D, which are robust under misspecification (robust SE).
(4) The test value for the "Size" columns is 1, the true value of β_4; the test value for the "Power" columns is 0.
(5) See Section 4 for the details of the partially linear model example and the Monte Carlo experiments.

underestimates the randomness in the averaging estimators $\hat{\beta}_{n,\hat{w}_n,2}$, $\hat{\beta}_{n,\hat{w}_n,3}$ and $\hat{\beta}_{n,\hat{w}_n,4}$, since it treats the averaging weight $\hat{w}_n$ as non-random.

Tables D.1 - D.3 report for different ρ values the rejection rates of $\hat{\beta}_{n,SP}$ with the common standard error and those of $\hat{\beta}_{n,\hat{w}_n}$ with both the naive and the two-step inference methods ($S = 1000$ random draws in the second step). Table D.1 is for β_2, Table D.2 is for β_3 and Table D.3 is for β_4. Two variations of the naive inference methods for $\hat{\beta}_{n,\hat{w}_n}$ are considered. The "Naive" one uses the common estimators of $V_{F,P}$ and C_F when computing the standard error, but they can be biased under misspecification (recall the discussion after (C.3)). The "Naive (robust SE)" one uses the robust influence functions (C.2) - (C.3) when computing the standard error. For the "Size" columns, the test value is the true value of the coordinate (i.e., 3 for β_2, 2 for β_3 and 1 for β_4); for the "Power" columns, the test value is 0. Table D.1 - D.3 also report the average ratios between the lengths of the two-step confidence intervals of $\hat{\beta}_{n,\hat{w}_n,j}$ and of the standard confidence intervals of $\hat{\beta}_{n,SP,j}$ $(j = 2, 3, 4)$.

All these results are categorically similar to those for β_1.

Appendix E Justification for $V_{SP} \geq V_P$ in Condition 2

Justification for Condition 2(i)

This section provides rationale of $V_{SP} \geq V_P$ in Condition 2(i) based on the semiparametric efficiency theory and Le Cam's third lemma. (The subscript F is suppressed throughout this subsection for notational simplicity.) What follows is not the proof of Condition 2(i), since Condition 2 is a maintained assumption and can be verified with corresponding primitive conditions for a specific model (like Appendix D for the partially linear model). This subsection merely argues that $V_{SP} \geq V_P$ in Condition 2(i) holds for quite general semiparametric models as it does not require much more than the setup of the semiparametric model.

Consider a set $\mathcal{P}$ consisting of densities $f(z|\beta, \beta_P, h, \eta)$, where h is the nuisance parameter identified by the objective function $R(h)$ in (3.2), β is the parameter of interest identified by h and the objective function $Q(\beta, h)$ in (3.1), β_P is the parameter identified by g_γ and the objective function $Q(\beta, g_\gamma)$ in (3.8),[11] and let $\eta \in \mathcal{E}$ denote the parameter that determines the features of the distribution of Z other than those characterized by β, β_P and h.[12] Maintain the assumption that the true density is in $\mathcal{P}$; in other words, $\mathcal{P}$ is the semiparametric model. Let $\underline{V_{SP}}$ and $\underline{V_P}$ denote the efficiency bounds of β and β_P, respectively.

Let $\delta \equiv \beta_P - \beta$, then the densities in $\mathcal{P}$ can be rewritten as $f(z|\beta, \beta+\delta, h, \eta)$. For any $f(z|\beta, \beta+\delta, h, \eta) \in \mathcal{P}$, one can define a parametric model (a subset of $\mathcal{P}$) that incorporates the parametric restriction

$$\begin{aligned}\mathcal{P}_{\beta,\delta,\gamma} \equiv \{ f(z|\beta, \beta+\delta, g_\gamma, \eta) : \ & \beta, \delta \in \mathbb{R}^k, \ \gamma \in \mathbb{R}^t; \\ & \gamma \text{ is identified by the objective function } R(g_\gamma) \text{ in (3.7)}; \\ & \delta = 0 \text{ only if } h = g_\gamma \text{ for some } \gamma \in \mathbb{R}^t \big\}.\end{aligned}$$

Note that this parametric model internalizes the parametric restriction and Condition 1(i) that the parametric restriction leads to bias if misspecified. Also note that $\mathcal{P}_{\beta,\delta,\gamma}$ may or may not include $f(z|\beta, \beta+\delta, h, \eta)$ itself, depending on whether h admits the parametric restriction g_γ.

If the density $f(z|\beta, \beta+\delta, h, \eta)$ itself belongs to $\mathcal{P}_{\beta,\delta,\gamma}$ (i.e., $h = g_\gamma$ for some $\gamma \in \mathbb{R}^t$), then the parametric restriction is correctly specified, and $\mathcal{P}_{\beta,\delta,\gamma}$ is a parametric submodel – that is, a parametric model that

[11] Recall that g_γ is a given parametric function characterized by $\gamma \in \mathbb{R}^t$, which is identified by the objective function $R(g_\gamma)$ in (3.7). In fact, for given g_γ function, one can rewrite $f(z|\beta, \beta_P, h, \eta)$ as $f(z|\beta, \beta_P, h, g_\gamma, \eta)$ to make the dependence of β_P on γ explicit, but g_γ is suppressed here for notational simplicity.

[12] This follows the setup in the proof of Lemma 1 in Ackerberg et al. (2014). η may have infinite dimension.

includes the true DGP – like that defined by Bickel, Klaassen, Ritov and Wellner (1993, Definition 1 on page 46) or Tsiatis (2006, page 59). As a result, one has $\underline{V_{SP}} \geq \underline{V_P}$ by the definition of the semiparametric efficiency bound – that is, the efficiency bound of the semiparametric model is the supremum of efficiency bounds of all parametric submodels – such as equation (2) on page 46 in Bickel et al. (1993) or equation (4.16) in Tsiatis (2006). At the same time, the construction of $\mathcal{P}_{\beta,\delta,\gamma}$ dictates that $\delta = 0$ and $\beta = \beta_P$ when $f(z|\beta, \beta+\delta, h, \eta) \in \mathcal{P}_{\beta,\delta,\gamma}$. This implies that the statement $V_{SP} \geq V_P$ in Condition 2(i) is a plausible condition when the parametric restriction is correctly specified.

Remark 3. *By the definition of the efficiency bounds, one has $V_{SP} \geq \underline{V_{SP}}$ and $V_P \geq \underline{V_P}$, and either equality holds if the corresponding estimator is efficient. Because the above shows that $\underline{V_{SP}} \geq \underline{V_P}$, the statement $V_{SP} \geq V_P$ in Condition 2(i) only means that $\hat{\beta}_{n,P}$ is at least as efficient as $\hat{\beta}_{n,SP}$, but does not require $\hat{\beta}_{n,P}$ to be efficient in general. For instance, if $\underline{V_{SP}} > \underline{V_P}$ or $V_{SP} > \underline{V_{SP}}$, then there is room between V_{SP} and $\underline{V_P}$ such that it is possible that the asymptotic variance V_P of some inefficient parametric estimator $\hat{\beta}_{n,P}$ satisfies $V_{SP} > V_P$.*[13]

What follows shows that the relationship $V_{SP} \geq V_P$ remains invariant if the parametric restriction deviates from correct specification to mild misspecification. For any fixed density $f(z|, \beta^*, \beta^*, g_{\gamma^*}, \eta^*)$ in $\mathcal{P}_{\beta^*,\delta^*,\gamma^*}$ (i.e., $\delta^* = 0$ by the construction of $\mathcal{P}_{\beta^*,\delta^*,\gamma^*}$), let P denote the resulting probability measure, and let $P_n = P$ be a sequence of such probability measures (same for all $n \in \mathbb{N}$). Note that P_n corresponds to the case where the parametric restriction is correctly specified. For any nuisance function h that does not admit the functional form g_γ, one has $f(z|\beta^*, \beta^*+\delta, h, \eta^*) \notin \mathcal{P}_{\beta^*,\delta^*,\gamma^*}$ and $\delta \neq 0$ by the construction of $\mathcal{P}_{\beta^*,\delta^*,\gamma^*}$. Inspired by Theorem 7.2 in Van der Vaart (2000), consider a sequence of such nuisance functions, denoted by h_n, such that the resulting densities $f(z|\beta^*, \beta^*+\delta_n, h_n, \eta^*)$ satisfy $\delta_n = \frac{d_n}{\sqrt{n}}$, $d_n \to d$ for some $d \in \mathbb{R}^k$ with $\|d\| \in (0, \infty)$ and the corresponding g_{γ_n} are g_{γ^*}. Note that the sequence of h_n, by the construction of $\mathcal{P}_{\beta^*,\delta^*,\gamma^*}$, converges to g_{γ^*}, since the corresponding δ_n converges to zero. Let Q_n denote the resulting sequence of probability measures, and it corresponds to the case where the parametric restriction is mildly misspecified. Under a technical condition called *differentiable in quadratic mean at β^**,[14] the log likelihood ratio between Q_n and P_n admits the following Taylor expansion with respect to β_P (i.e., $\beta^* + \delta_n$) around β^*:

$$\begin{aligned}\log \prod_{i=1}^n \frac{dQ_n}{dP_n} &= \log \prod_{i=1}^n \frac{f_n(Z_i|\beta^*, \beta^*+\delta_n, h_n, \eta^*)}{f(Z_i|\beta^*, \beta^*, g_{\gamma^*}, \eta^*)} \\ &= d' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\beta_P}(Z_i) \right) - \frac{1}{2} d' \left(-\frac{1}{n} \sum_{i=1}^n \ddot{\ell}_{\beta_P}(Z_i) \right) d + o_p(1),\end{aligned}$$

where $\dot{\ell}_{\beta_P}(z) \equiv \frac{\partial f(z|\beta^*,\beta^*,g_{\gamma^*},\eta^*)/\partial \beta_P}{f(z|\beta^*,\beta^*,g_{\gamma^*},\eta^*)}$ is the score function with respect to β_P under P_n evaluated at β^*, and $\ddot{\ell}_{\beta_P}(z) \equiv \frac{\partial^2 f(z|\beta^*,\beta^*,g_{\gamma^*},\eta^*)/\partial \beta \partial \beta'}{f(z|\beta^*,\beta^*,g_{\gamma^*},\eta^*)}$ is the corresponding Hessian matrix. Note that $\mathbb{E}_{P_n}[\dot{\ell}_{\beta_P}(Z_i)] = 0$ and $\mathbb{E}_{P_n}[-\ddot{\ell}_{\beta_P}(Z_i)] = \mathcal{I}_{\beta_P}$ (the Fisher information matrix with respect to β_P). By the central limit theorem and

[13] A well known special case is the inverse probability weighted (IPW) estimator of the average treatment effect (ATE) with series logit propensity score. For the ATE, Hahn (1998) proves that $\underline{V_{SP}} = \underline{V_P}$ under the correct specification of the parametric restriction, and Hirano, Imbens and Ridder (2003) show that the IPW estimator with series logit propensity score satisfies $V_{SP} = \underline{V_{SP}}$. Together, these require that the parametric estimator has to be efficiency for $V_{SP} \geq V_P$ to hold (with equality). The author thanks an anonymous referee for pointing this out.

[14] See (7.1) in Van der Vaart (2000), for example. This assumption is common and maintained for the majority of the models in the M estimation literature.

Cramér-Wold theorem, it can be shown that

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_{n,SP} - \beta) \\ \sqrt{n}(\hat{\beta}_{n,P} - \beta) \\ \log \prod_{i=1}^{n} \frac{dQ_n}{dP_n} \end{pmatrix} \xrightarrow{P_n} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ -\frac{1}{2} d' \mathcal{I}_{\beta_P} d \end{pmatrix}, \begin{pmatrix} V_{SP} & C & \tau_{SP} \\ C & V_P & \tau_P \\ \tau'_{SP} & \tau'_P & d' \mathcal{I}_{\beta_P} d \end{pmatrix} \right), \tag{E.1}$$

where the symbol $\xrightarrow{P_n}$ means that the left hand side converges in distribution to the right hand side if P_n is the true distribution of the data.[15] This fulfills the assumption of Le Cam's third lemma (Example 6.7 in Van der Vaart, 2000), so by this lemma one gets

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_{n,SP} - \beta) \\ \sqrt{n}(\hat{\beta}_{n,P} - \beta) \end{pmatrix} \xrightarrow{Q_n} \mathcal{N}\left(\begin{pmatrix} \tau_{SP} \\ \tau_P \end{pmatrix}, \begin{pmatrix} V_{SP} & C \\ C & V_P \end{pmatrix} \right). \tag{E.2}$$

That is, Le Cam's third lemma implies that the asymptotic variance-covariance matrices of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ remain invariant whether the parametric restriction is correctly specified or mildly misspecified. Together with the earlier condition that $V_{SP} \geq V_P$ under the correct specification, it provides the rationale behind $V_{SP} \geq V_P$ in Condition 2(i).

Remark 4. *In* (E.1) *and* (E.2), $\tau_{SP} \equiv \mathbb{E}_{P_n}[\psi_{SP}(Z_i)\dot{\ell}_{\beta_P}(Z_i)]d$ *and* $\tau_P \equiv \mathbb{E}_{P_n}[\psi_P(Z_i)\dot{\ell}_{\beta_P}(Z_i)]d$ *by the central limit theorem, where* $\psi_{SP}(z)$ *and* $\psi_P(z)$ *are the (centered) influence functions of* $\hat{\beta}_{n,SP}$ *and* $\hat{\beta}_{n,P}$*, respectively. Note that* $\hat{\beta}_{n,SP}$ *is a regular and asymptotically linear (RAL) estimator of* β *but* $\hat{\beta}_{n,P}$ *is an RAL estimator of* β_P*,*[16] *then by Theorem 4.2 in Tsiatis (2006), one has* $\mathbb{E}_{P_n}[\psi_{SP}(Z_i)\dot{\ell}_{\beta_P}(Z_i)] = 0$ *and* $\mathbb{E}_{P_n}[\psi_P(Z_i)\dot{\ell}_{\beta_P}(Z_i)] = I_k$ *(i.e., the* $k \times k$ *identity matrix), further implying that* $\tau_{SP} = 0$ *and* $\tau_P = d$*. This, combined with the above argument for* $V_{SP} \geq V_P$*, indicates that the joint asymptotic distribution postulated in Condition 2(i) is in fact a general result for the semiparametric model and the estimators considered in this paper.*

Justification for Condition 2(ii)

Note that the asymptotic properties of the semiparametric estimator $\hat{\beta}_{n,SP}$ do not depend on whether $\|d\| < \infty$ or $\|d\| = \infty$, so one still has $n^{1/2}(\hat{\beta}_{n,SP} - \beta_{F_n}) \xrightarrow{d.} \xi_{F,SP}$ under the same primitive conditions like those for Condition 2(i).

To study the asymptotic properties of the parametric estimator $\hat{\beta}_{n,P}$ when $\|d\| = \infty$, consider two cases: (i) $\delta_{F_n} = o(1)$; and (ii) $\|\delta_{F_n}\| > c$ for some $c > 0$. For case (i), let $\psi_{F,P}(z)$ denote the (centered) influence function of $\hat{\beta}_{n,P}$ under DGP F, which is an $O_p(1)$ term, then by the definition of $\beta_{F,P}$ and δ,

$$\begin{aligned} n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n,P}) &= n^{-1/2} \sum_{i=1}^{n} \psi_{F_n,P}(Z_i) + o_p(1) \\ \Longrightarrow n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n}) &= n^{1/2}\delta_{F_n} + O_p(1). \end{aligned} \tag{E.3}$$

Note that the presumption of Condition 2(ii) is that $\|n^{1/2}\delta_{F_n}\| \to \|d\| = \infty$, then $n\delta'_{F_n}\delta_{F_n} \to \infty$, which together with (E.3) implies that $\|n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n})\| \xrightarrow{p.} \infty$.

For case (ii), note that $\beta_{F,P}$ is identified in (3.8), then under the same conditions for $\hat{\beta}_{n,SP} = \beta_{F_n} + o_p(1)$, one gets $\hat{\beta}_{n,P} = \beta_{F_n,P} + o_p(1)$.[17] This, combined with the presumption that $\|\delta_{F_n}\| = \|\beta_{F_n,P} - \beta_{F_n}\| > c$,

[15] The symbol $\xrightarrow{Q_n}$ below is understood in a similar way.

[16] Recall that this paper maintains the assumption that $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are both locally regular estimators, so that their asymptotic properties hinge on the influence functions.

[17] This is a familiar result for pseudo-true parameter values (e.g., Newey and McFadden, 1994, Section 2).

implies that

$$\|n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n})\| \geq |\|n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n,P})\| - \|n^{1/2}\delta_{F_n}\|| = \|n^{1/2}\delta_{F_n}\| \cdot (1 + o_p(1)) \xrightarrow{p.} \infty.$$

References

Ackerberg, D., X. Chen, and J. Hahn (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics* 94(2), 481-498.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014). Asymptotic efficiency of semiparametric two-step GMM. *Review of Economic Studies* 81(3), 919-943.

Bickel, P. J., C. A. Klaassen, J. Ritov, and J. A. Wellner (1993). Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press Baltimore.

Donald, S. G. and W. K. Newey (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis* 50(1), 30-40.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2), 135-331.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161-1189.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric M-estimators. *Journal of Econometrics* 159(2), 252-266.

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

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothetis testing. *Handbook of Econometrics* 4, 2111-2245.

Tsiatis, A. (2006). Semiparametric theory and missing data. Springer Science & Business Media.

Van der Vaart, A. W. (2000). Asymptotic statistics. *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.