*Econometric Theory*, **40**, 2024, 652–687.
doi:10.1017/S0266466622000548

# AN AVERAGING ESTIMATOR FOR TWO-STEP M-ESTIMATION IN SEMIPARAMETRIC MODELS

RUOYAO SHI
*University of California, Riverside*

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 a 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 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.

## 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

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 was generously supported by UC Riverside Regents' Faculty Fellowship 2019–2020. Zhuozhen Zhao provided great research assistance. All remaining errors are the author's. Address correspondence to Ruoyao Shi, Department of Economics, University of California, Riverside, Riverside, CA, USA; e-mail: ruoyao.shi@ucr.edu.

[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).

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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 $\beta$ denote the unknown parameter of interest, and let $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ denote the semiparametric and 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). 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 lower bounds of the finite sample truncated quadratic risk difference between the averaging estimator and the semiparametric estimator over a large class of data generating processes (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 includes those under which the parametric restrictions are correctly specified, mildly misspecified and severely misspecified, the uniform

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

dominance result here asserts that the 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 a two-step M-estimation framework for semiparametric models as it encompasses maximum likelihood estimator (MLE), GMM, many kernel-based and sieve estimators, and so forth, as well as regular one-step M-estimators, as special cases.

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 inflict 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 (see also, 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.

## 1.1. 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

[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.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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-squares 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) and DiTraglia (2016) consider averaging GMM estimators, and Kitagawa and Muris (2016) analyze averaging semiparametric estimators of the treatment effects on the treated 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 (2022). 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. This paper differs from this 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 with a uniform asymptotic approach, instead of the pointwise local asymptotic approach (Le Cam, 1972; Van der Vaart, 2000, Chap. 7) often taken in the literature. Finally, the Stein-type condition for the uniform dominance of the averaging estimator in this paper is stronger than 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

[4] With a certain choice of the weighting matrix in the loss function, the main theorem of this paper (Theorem 1) requires the dimension of the parameters of interest to be at least 4. 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 (2020).

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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 Heckman's (1979) two-step 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 et al. (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 and Robinson (1989) using kernel estimators. 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's estimator (e.g., Van der Vaart, 2000, Exam. 8.1), among others (Leeb and Pötscher, 2005; Yang, 2005; Leeb and Pötscher, 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.

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, whereas 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 et al., 2018, 2022) removes impacts of the nuisance function estimation bias

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

(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 (2021) 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.

### 1.2. 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. The Appendix gives the proofs of the results in Section 3. Appendix B provides the proofs of the intermediate lemmas in the Appendix of this article. Appendix C presents an alternative method of computing the averaging weight. Appendix D details additional theoretical and MC results for the example in Section 4. Appendix E discusses the justification for the high-level Condition 2. Appendixes B–E are available in the Supplementary Material associated with this article.

## 2. COMPUTING THE AVERAGING ESTIMATOR

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.

### 2.1. Averaging Weight

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 $\beta$ and a chosen symmetric positive semidefinite weighting matrix $\Upsilon$,[6] define the loss function to

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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 $\Upsilon$ is chosen by the researcher and reflects how much the researcher values the estimation accuracy of each coordinate of $\beta$. If the researcher treats every coordinate equally, then they 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 $\beta$, then they may choose $\Upsilon$ 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 $\Upsilon$ here affords more flexibility. Note that both the loss function and the averaging weight (to be introduced later) depend on $\Upsilon$, 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

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

[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).

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

where tr[$\cdot$] indicates the trace of a square matrix.[8] This weight falls in the interval $[0,1]$ with probability one, and the reason is detailed in Appendix C in the Supplementary Material.[9]

**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

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

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 $\mathrm{tr}[\Upsilon(\widehat{V}_{F,SP} - \widehat{V}_{F,P})]$, 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 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's (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).

## 2.2. Bootstrapping Asymptotic Variance–Covariance Matrices

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 $\left(\hat{\beta}'_{n,SP}, \hat{\beta}'_{n,P}\right)'$.[10]

Because the consistency of the bootstrap distribution does not guarantee the consistency of the bootstrap second moment (Hahn and Liao, 2021), one needs

[8] 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 $\Upsilon$ and properties of the trace operator. The same goes for $C_F$ and $C'_F$.

[9] 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.

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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 $j$th 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 $\nu$ and small $c_0$, define $T_j \equiv \max\{\nu|\hat{\beta}_j|, c_0\}$ for each coordinate.[11] For all $b$ and all $j$, define

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

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 in the Supplementary Material 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.

### 3.1. Uniform Dominance

Suppose $\beta_F$, 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

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

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

[11] The validity of Shao's (1992) method does not rely on any specific values of $\nu$ 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 MC experiments in Section 4 of this paper.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

argument, the presence of $h$ and how it is modeled generally affect the asymptotic properties of $\beta$ estimators through $Q_F$, even in the absence of estimation error of $h$.[12] 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 $\beta$ 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 stated in (1.1), this paper considers averaging two common two-step M-estimators of $\beta$: 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 by $\mathcal{H}_g$, where $g$ is a function that is known up to a finite-dimensional vector of unknown parameters $\gamma$. 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

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

[12] 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).

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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 $\gamma_F$ as the unique minimizer (assume it exists) of the following problem:

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

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 $g_{\gamma_F}$ induced by the pseudo-true parameter $\gamma_F$ is different from the true nuisance function $h_F$ identified in (3.2). In consequence, $\beta_{F,P}$ in general will be different from $\beta_F$, the true parameter of interest identified in (3.1). Let $\delta_F$ 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 $\beta$ and an arbitrary real number $\zeta$, 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 $\zeta$ could be arbitrarily large. The bounds of the truncated risk differences for finite sample size $n$ are defined as

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

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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

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

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

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 $\delta_F$ 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 $\beta_F$, which rules out the uninteresting special case that $\beta_F$ may be consistently estimable even with severely misspecified parametric restrictions. As a result, the degree of misspecification can be indexed by $\delta_F$, 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 $F$,

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

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

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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, Def. 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, Sect. 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 $\delta_{F_n}$ and the presumption $n^{1/2}\delta_{F_n} \to d$. Moreover, $V_{F,SP} \geq V_{F,P}$ states the intuition that imposing parametric restrictions generally leads to (weak) efficiency gain.[13] Formal justification of Condition 2(i) is in Appendix E in the Supplementary Material, 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, Exam. 6.7) and the definition of semiparametric efficiency bound (see Bickel et al., 1993, Chap. 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 $\mathcal{H}$),[14] 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.[15] Second, when $\|d\| < \infty$ but $\|d\| \neq 0$, the

[13] 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 in the Supplementary Material.

[14] That is, the difference between the two is a negative semidefinite matrix.

[15] Following the Ackerberg et al. (2014) approach, one needs to define another nuisance parameter $\eta$, which captures the features of the distribution of data $Z$ other than those determined by $\beta$ and $h$, then characterize the tangent space

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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 in the Supplementary Material.

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, Sect. 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*

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

*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 the Appendix. □

(see Newey, 1990; Bickel et al., 1993) for both the unrestricted and the restricted models. The efficient score function of $\beta$ in each model is therefore the projection residual of the score function of $\beta$ 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 $\beta$ 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.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

**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

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

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(i) still hold. When $V_{F,SP} < V_{F,P}$, on the other hand, it is easy to see that the weight $\tilde{w}_n$ converges to zero in probability, so the results in Lemma 1(ii) hold regardless of $\|d\|$ value. 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.

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 $\delta_F$ 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 $S(\tilde{F})$ and $S(F)$.[16] 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 in the Supplementary Material details the primitive conditions of Condition 2

[16] 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.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

for the partially linear model in Section 4 and verifies Conditions 1–3 for the parameterization used in the MC experiments in that section.

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

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

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 $\delta_F$ 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 $\delta$ ranges from being zero to approaching to zero at any rate that is not slower 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 $\delta$ explodes, is a constant, or approaches to zero at slower 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 prove 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.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

LEMMA 2. *Suppose: (i) Conditions 1–3 hold; and (ii) $tr(A_F) > 0$ and $tr(B_F) > 0$, where $A_F$ and $B_F$ are defined in (3.20).*[17] *Then*

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

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

**Proof.** See the Appendix. □

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 the Appendix. □

To give some intuition for the conditions of the uniform 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 $V_{F,SP} - C_F$ require 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),

[17] As shown in the Appendix (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 $\text{tr}(A_F) > 0$.

and 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

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

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

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

Furthermore, 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) converges 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,[18] and $w_F$ in (3.21) is a random variable and in general not unbiased for $w^*$ in light of 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 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. The subsequence technique in Cheng et al. (2019), therefore, is necessary for proving Theorem 1.

## 3.2. Inference

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

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

To adapt Claeskens and Hjort's (2008) 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 distribution 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 by $CI_{1-\alpha_2}(\Delta_n|d,\widehat{V})$, 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 $\alpha_2/2$ and $1-\alpha_2/2$ quantiles of the simulated $\bar{\xi}_{F,d}$ values given in (3.22).[19]

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 $1-\alpha_1$:

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

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 $\chi^2$ distribution with degrees of freedom $k$.

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

(1) For any confidence level $1-\alpha$, pick $\alpha_1$ and $\alpha_2$ 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 $\Delta_n$ 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 $\alpha_1$ and $\alpha_2$ such that $\alpha_1+\alpha_2=\alpha$, the $1-\alpha$ confidence set of $\beta_F$ 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 then 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 $\beta_F$ based on the data.

The next lemma follows Claeskens and Hjort (2008, Sect. 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 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.

[19] 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 $\widehat{C}_n$.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

LEMMA 3. *Suppose Conditions 1–3 hold. Let $\alpha_1$ and $\alpha_2$ be fixed nonnegative 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 the Appendix. □

In contrast to this two-step method, a naive inference method based on the averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ would treat the averaging weight $\hat{w}_n$ as nonrandom and compute the asymptotic variance–covariance matrix of the averaging estimator 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

### 4.1. Example: Partially Linear Model

One is interested in estimating $\beta$ in a partially linear model

$$Y = X_1'\beta + s(X_1,X_2) + U, \tag{4.1}$$

where $\mathbb{E}(U|X_1,X_2) = 0$, $X_1$ is a $k\times 1$ vector, $X_2$ is an $l\times 1$ vector, and $X_1$ and $X_2$ are assumed not to overlap for simplicity. The identification of $\beta$ requires that $s(X_1,X_2)$ and $X_1$ are not perfectly collinear.[20]

The estimator of $\beta$ 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 $\hat{\beta}_{n,SP}$.[21] 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 $\beta$ could serve as the parametric estimator $\hat{\beta}_{n,P}$.[22]

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

[20] To be precise, $\mathbb{E}\{[X_1 - \mathbb{E}[X_1|s(X_1,X_2)]\cdot[X_1 - \mathbb{E}[X_1|s(X_1,X_2)]'\}$ is positive definite.

[21] Many semiparametric estimators of $\beta$ 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 the discussion on pages 5–8 of Ahn and Powell, 1993, and the references therein), many semiparametric estimators of $\beta$ in sample selection models (with potentially nonparametric selection equation) have been proposed and examined under various identification conditions, such as Gallant and Nychka (1987), Newey, Powell, and Walker (1990), Ahn and Powell (1993), and Newey (2009). They could all serve as the $\hat{\beta}_{n,SP}$ in this paper, provided that the conditions in Section 3 are satisfied.

[22] 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 $s(x_1,x_2)$.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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.[23] Third, the Stein-type condition amounts to a dimensionality condition ($k \geq 4$) for $\Upsilon = \left(V_{F,SP} - V_{F,P}\right)^{-1}$ (discussed after Theorem 1) and it can be easily fulfilled in this example.

## 4.2. Monte Carlo Experiments

In the MC 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 $j$th 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:

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

whereas 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).[24] 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

$$V_X = \begin{bmatrix} 0.5^2 \times I_4 & 0.05 \times L_{4\times4} \\ 0.05 \times L_{4\times4} & 0.5^2 \times I_4 \end{bmatrix}, \tag{4.4}$$

where $\ell_8$ is an $8 \times 1$ vector of ones, $I_4$ is the $4 \times 4$ identity matrix, and $L_{4\times4}$ is a $4 \times 4$ 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 $\rho$, 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

[23] Appendix D in the Supplementary Material 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.

[24] 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 MC experiments.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

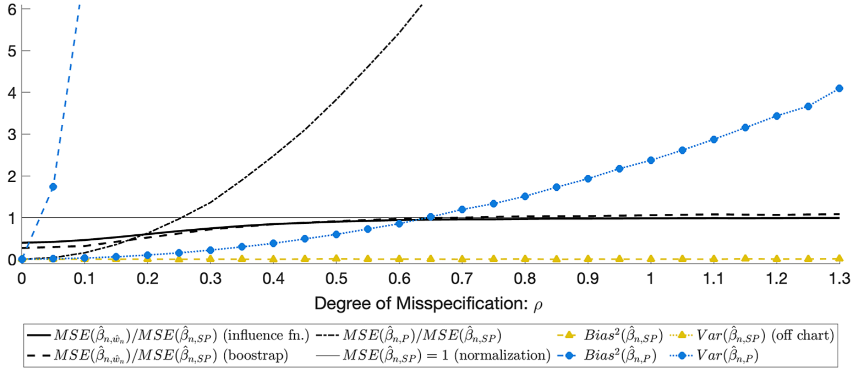


**Figure 1.** MC MSE, bias$^2$, and variance of $\hat\beta_{n,SP}$, $\hat\beta_{n,P}$, and $\hat\beta_{n,\hat w_n}$ for the partially linear model. *Notes*: (1) Bootstrap results are based on $R = 1{,}000$ MC replications, $n = 1{,}000$ sample size, and $B = 200$ bootstrap replications. (2) All other results are based on $R = 10{,}000$ MC replications and $n = 1{,}000$ 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 MC experiments.

is $n = 1{,}000$, and the number of MC replications is $R = 10{,}000$.[25] 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 $\rho$ 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 MSEs of an estimator being below this unity benchmark means that this estimator 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 in the Supplementary Material) are represented by the thick solid black line, whereas those using the bootstrapped asymptotic variance–covariance matrix estimates ($B = 200$ bootstrap replications) are represented by the dashed black line.[26] 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.[27] 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 $\rho$ and shoots outside the figure range before $\rho$ reaches 0.1. The variances of $\hat\beta_{n,SP}$ are represented by the dotted yellow line with triangle markers, and those of $\hat\beta_{n,P}$

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

[26] To save time, the bootstrap averaging estimator is only computed for $R = 1{,}000$ replications with 0.1 step width of $\rho$ values.

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

by the dotted blue line with round markers; the former remains stable around the level of 45 and is outside the figure range.

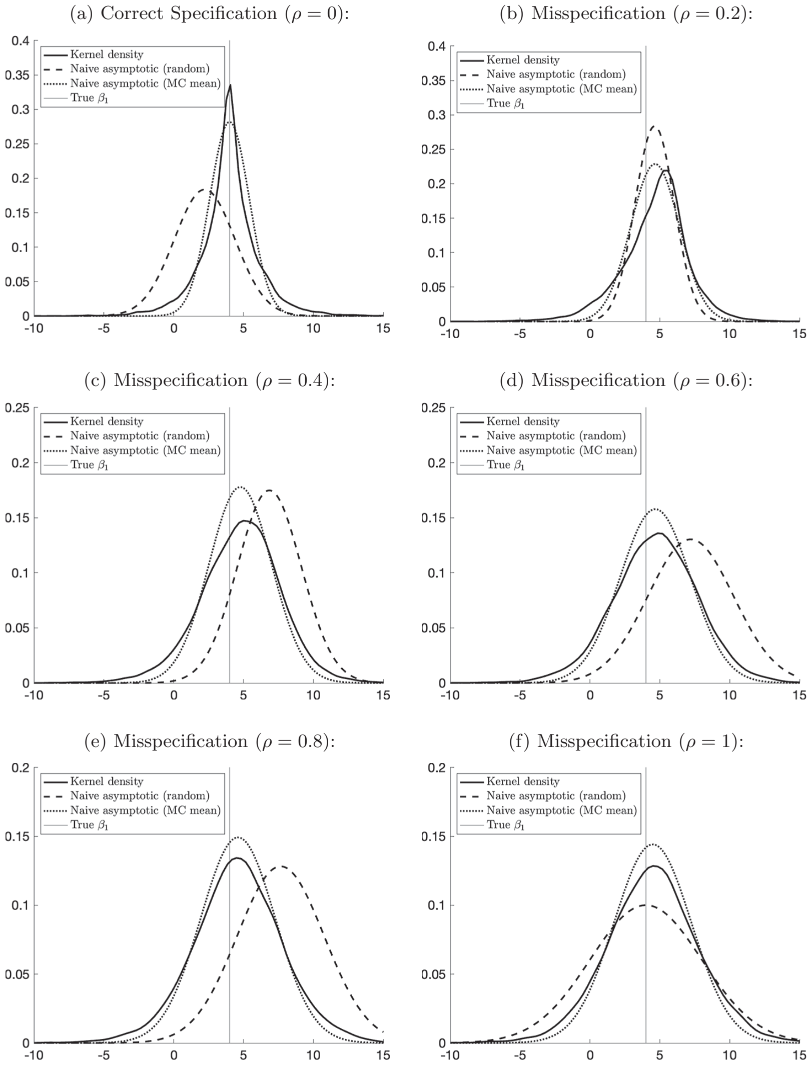
Figure 2 plots the MC distributions (kernel densities) of the first coordinate of the averaging estimator $\hat{\beta}_{n,\hat{w}_n}$ (thick solid lines) for representative $\rho$ values. In the same figures, the normal distributions based on the naive inference method with the common standard error are represented by the thick dashed lines (one randomly chosen MC replication) and dotted lines (averaged over all MC replications). It is obvious that the naive inference method miscalculates the randomness in the averaging estimator $\hat{\beta}_{n,\hat{w}_n,1}$, since it treats the averaging weight $\hat{w}_n$ as nonrandom.

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

Table 1 reports for different $\rho$ 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 ($S = 1,000$ random draws in the second step) for $\beta_1$, the first coordinate of $\beta$. 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 (D.2) when computing the standard error (see Appendix D in the Supplementary Material for details). For the "Size" columns, the test value is 4, the true value of $\beta_1$; 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 MC results. First, regardless of the degree of misspecification, $\hat{\beta}_{n,SP}$ has almost zero bias but very large and stable variance, whereas $\hat{\beta}_{n,P}$ has much smaller variance but rapidly increasing bias. Second 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 $\rho$ increases beyond 0.6. Third, 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. Fourth, 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 extent. Fifth, the asymptotic distributions based on the naive inference method in Figure 2 badly approximate the actual MC distributions of the averaging estimator $\hat{\beta}_{n,\hat{w}_n,1}$ for all $\rho$ values. Sixth, 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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press



(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 ($\rho = 1$):

**Figure 2.** True versus naive distributions of $\hat{\beta}_{n,\hat{w}_n,1}$ for the partially linear model. *Notes*: (1) All distributions are based on $R = 10{,}000$ MC replications and $n = 1{,}000$ sample size. (2) The solid lines represent the MC distributions of $\hat{\beta}_{n,\hat{w}_n,1}$, the averaging estimator of $\beta_1$. The 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, whereas the latter show the average over all MC replications. (3) See Section 4 for the details of the partially linear model example and the MC experiments.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

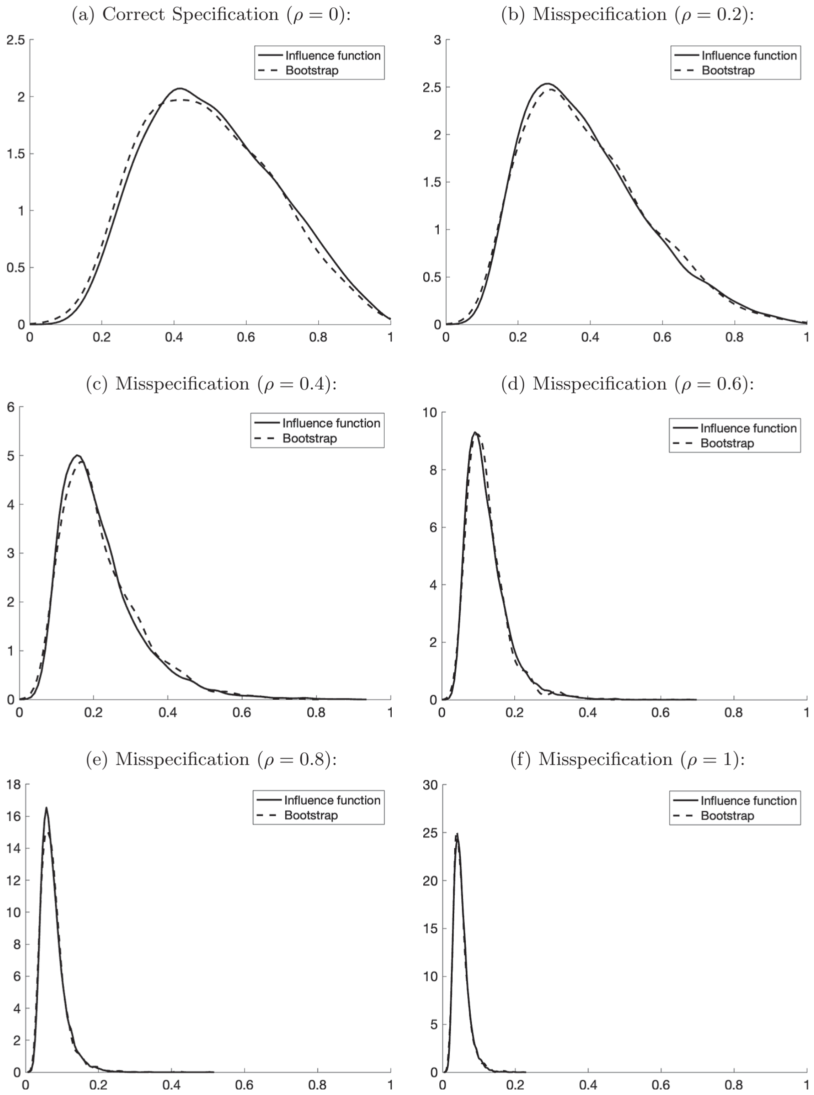


**FIGURE 3.** Distributions of $\hat{w}_n$ for the partially linear model. *Notes:* (1) Bootstrap results are based on $R = 1,000$ MC replications, $n = 1,000$ sample size, and $B = 200$ bootstrap replications. (2) All other results are based on $R = 10,000$ MC replications and $n = 1,000$ sample size. (3) The distributions of the averaging weight $\hat{w}_n$ concentrate toward zero as $\rho$, the degree of misspecification, increases. (4) See Section 4 for the details of the partially linear model example and the MC experiments.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

**TABLE 1.** Rejection rates for $\hat{\beta}_{n,\hat{w}_n,1}$ in the partially linear model (5% level)

| $\rho$ | $\hat{\beta}_{n,SP,1}$ | | $\hat{\beta}_{n,\hat{w}_n,1}$ | | | | | | 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 $\beta_1$. The results for the other three coordinates are reported in Tables D.1–D.3 of Appendix D in the Supplementary Material.

(2) All results are based on $R = 10{,}000$ MC replications and $n = 1{,}000$ sample size. The two-step inference method uses $S = 1{,}000$ random draws 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 nonrandom, 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)); and the second (robust SE) uses the robust influence function (D.2) when computing the standard error (see Appendix D for details).

(4) The test value for the "Size" columns is 4, the true value of $\beta_1$; the test value for the "Power" columns is 0.

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

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 MC results for $\beta_1$, the first coordinate of $\beta$. Similar results for the other three coordinates are reported in Figures D.1–D.3 and Tables D.1–D.3 in Appendix D in the Supplementary Material.

## 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 a 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.

## APPENDIX. Proofs of the Theorems

### A.1. Proof of Lemma 1

**Proof.** Part (i). Note that in this case, $\widehat{V}_{n,SP}$, $\widehat{V}_{n,P}$, and $\widehat{C}_n$ are consistent estimators of $V_{F,SP}$, $V_{F,P}$, and $C_F$, respectively, then the result 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 $\kappa$ 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 $\zeta$, define

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

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

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

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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

Note that $r(u_{F,d})$ in (A.5) is just what is defined in (3.27). For any positive finite $\zeta$, 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–A.4 can be found in Appendix B in the Supplementary Material.

LEMMA A.1. *Suppose Conditions 1–3 hold. Then,*

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

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

LEMMA A.2. *Suppose Conditions 1–3 hold. Then,*

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

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

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

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

## A.2. 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[28]

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

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). □

## A.3. Proof of Theorem 1

**Proof.** By Lemma 2, it suffices to show that $\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}) \le 0$ 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

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

[28] This is because Lemma A.4 means that for $\forall\epsilon > 0$, there exists a large enough number $C$ such that for all $\zeta \ge C$ one has $\sup_{u_{F,d}\in\mathcal{U}} \left|r_{\zeta}(u_{F,d}) - r(u_{F,d})\right| < \epsilon$. This implies that for $\zeta \ge C$ 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 $\epsilon$ does not vary with $u_{F,d}$, so for $\zeta \ge C$, 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 $\zeta \ge C$, 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.

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)),

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

where

$$J_{1,F} \equiv \mathbb{E}\left[\frac{(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon\xi_{F,SP}}{\mathrm{tr}(B_F)+(\xi_{F,P}+d-\xi_{F,SP})'\Upsilon(\xi_{F,P}+d-\xi_{F,SP})}\right],$$

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

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 rewritten as

$$J_{1,F} = \mathbb{E}\left[\frac{(\tilde{\xi}_F+\tilde{d})'E(\tilde{\xi}_F+\tilde{d})}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right],$$

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

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:

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

Its derivative (transposed) is then

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

Note that $J_{1,F} = \mathbb{E}[\eta_F(\tilde{\xi}_F+\tilde{d})'E(\tilde{\xi}_F+\tilde{d})]$ and $\mathrm{tr}(E\tilde{V}_F) = -\mathrm{tr}[\Upsilon(V_{F,SP}-C_F)] = -\mathrm{tr}(A_F)$, 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

$$\begin{aligned}
J_{1,F} &= \mathbb{E}\left[\mathrm{tr}\left(\frac{\partial}{\partial x}\eta_F(\tilde{\xi}_F+\tilde{d})'E\tilde{V}_F\right)\right] \\
&= \mathbb{E}\left[\frac{-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right] - 2\mathbb{E}\left[\frac{\mathrm{tr}[D(\tilde{\xi}_F+\tilde{d})(\tilde{\xi}_F+\tilde{d})'E\tilde{V}_F]}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right] \\
&= \mathbb{E}\left[\frac{-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})}\right] + 2\mathbb{E}\left[\frac{-(\tilde{\xi}_F+\tilde{d})'E\tilde{V}_F D(\tilde{\xi}_F+\tilde{d})}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d})'D(\tilde{\xi}_F+\tilde{d})]^2}\right].
\end{aligned} \tag{A.22}$$

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

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

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

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

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

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

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

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

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 nonpositive. If, in addition, $\mathrm{tr}(A_F)\geq 4\rho_{\max}(A_F)$, then the first term in (A.26) will be nonpositive. 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.

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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. □

## A.4. Proof of Lemma 3

**Proof.** For any given $d$, one has

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

where the first equality holds by the way in which $CI_{1-\alpha_2}(\Delta_n|d, \widehat{V})$ is constructed, and 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 $\alpha_1 + \alpha_2 = \alpha$. □

## SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit: https://doi.org/10.1017/S0266466622000548.

## REFERENCES

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

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

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

Ahn, H. & 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, & 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, & P. Guggenberger (2020) Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics* 218(2), 496–531.

Andrews, D.W. & 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. & 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, & 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. & M. Kolesár (2021) Sensitivity analysis using approximate moment condition models. *Quantitative Economics* 12(1), 77–108.

Bang, H. & 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, & J.A. Wellner (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

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

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

Bonhomme, S. & M. Weidner (2021) Minimizing sensitivity to model misspecification. Preprint, arXiv:1807.02161.

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

Cao, W., A.A. Tsiatis, & 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, & I. Van Keilegom (2003) Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* 71(5), 1591–1608.

Cheng, X., Z. Liao, & 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, & 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, & J.M. Robins (2022) Locally robust semiparametric estimation. *Econometrica* 90(4), 1501–1535.

Claeskens, G. & N.L. Hjort (2008) *Model Selection and Model Averaging*. Cambridge University Press.

Crepon, B., F. Kramarz, & 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. & W.K. Newey (1994) Series estimation of semilinear models. *Journal of Multivariate Analysis* 50(1), 30–40.

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

Fessler, P. & 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, & M.T. Wells (2018) *Shrinkage Estimation*. Springer.

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

Hahn, J. & 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. & 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*, vol. 5, pp. 475–492. National Bureau of Economic Research.

Heckman, J.J. (1979) Sample selection bias as a specification error. *Econometrica* 47(1), 153–161.

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

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

Hjort, N.L. & 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* 60(3), 533–565.

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

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

Ichimura, H. & W. Newey (2017) The Influence Function of Semiparametric Estimators. CEMMAP Working paper CWP06/17, The Institute for Fiscal Studies, Department of Economics, University College London.

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

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

Judge, G.G. & 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. & 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. & K.I. Wolpin (1997) The career decisions of young men. *Journal of Political Economy* 105(3), 473–522.

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

Klein, R.W. & R.H. Spady (1993) An efficient semiparametric estimator for binary response models. *Econometrica* 61(2), 387–421.

Le Cam, L. (1972) Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 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. & B.M. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21(1), 21–59.

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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

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

Magnus, J.R., O. Powell, & 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. & 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. 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* 62(6), 1349–1382.

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

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

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

Newey, W.K. & J.L. Powell (1999) Two-Step Estimation, Optimal Moment Conditions, and Sample Selection Models. Working paper 99-06, Department of Economics, Massachusetts Institute of Technology.

Newey, W.K., J.L. Powell, & 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. In *Probability and Statsitics*, pp. 213–234. Wiley.

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

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

Peng, J. & Y. Yang (2022) On improvability of model selection by model averaging. *Journal of Econometrics* 229(2), 246–262.

Powell, J.L. (1986) Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* 54(6), 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, & J. Powell (eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, vol. 13, pp. 165–196. Cambridge University Press.

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. & 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. & 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 no. 5.

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press

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* 61(1), 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*, pp. 197–206. University of California Press.

Tsiatis, A.A., M. Davidian, & 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. & A.D. Woodland (1980) Sample selectivity and the estimation of labor supply functions. *International Economic Review* 21(2), 437–468.

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

https://doi.org/10.1017/S0266466622000548 Published online by Cambridge University Press