An Averaging Estimator for Two Step M Estimation in Semiparametric Models

Ruoyao Shi*

February, 2021

Abstract

In a two step extremum estimation (M estimation) framework with a finite dimensional parameter of interest and a potentially infinite dimensional first step nuisance parameter, I propose an averaging estimator that combines a semiparametric estimator based on nonparametric first step and a parametric estimator which imposes parametric restrictions on the first step. The averaging weight is the sample analog of an infeasible optimal weight that minimizes the asymptotic quadratic risk. I show that under mild conditions, the asymptotic lower bound of the truncated quadratic risk difference between the averaging estimator and the semiparametric estimator is strictly less than zero for a class of data generating processes (DGPs) that includes both correct specification and varied degrees of misspecification of the parametric restrictions, and the asymptotic upper bound is weakly less than zero.

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

*Department of Economics, University of California Riverside, ruoyao.shi@ucr.edu. The author thanks Colin Cameron, Xu Cheng, Denis Chetverikov, Yanqin Fan, Jinyong Hahn, Bo Honoré, Toru Kitagawa, Zhipeng Liao, Hyungsik Roger Moon, Whitney Newey, Geert Ridder, Aman Ullah, Haiqing Xu and the participants at various seminars and conferences for helpful comments. This project is generously supported by UC Riverside Regents' Faculty Fellowship 2019-2020. Zhuozhen Zhao provides great research assistance. All remaining errors are the author's.

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 in general less efficient than their parametric counterparts[1] which result from imposing certain parametric restrictions on the nonparametric components of semiparametric models. This efficiency defect of semiparametric estimators often render relatively imprecise estimates and low test power, especially when the parametric restrictions are correct or only mildly misspecified.

Recognizing such accuracy defect of semiparametric estimators, 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.

In this paper, I aim 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, I propose an averaging estimator that is a simple weighted average between the semiparametric estimator and the parametric estimator with a data-driven weight. Second, I prove that under mild conditions, the proposed averaging estimator exhibits (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.

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

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

where n is the sample size and $\hat{w}_n \in [0,1]$ is a data-driven weight elaborated in equation (3.8) below. Intuitively, the weight quantifies the asymptotic efficiency gain by imposing the parametric restrictions and the possible asymptotic misspecification bias by deviating

[1] In this paper, I 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 equation (3.5) below.

from the robust semiparametric benchmark. It then balances the two to reduce asymptotic quadratic risks compared to the semiparametric estimator.

I employ a *uniform* asymptotic theory to approximate the upper and the lower bounds of the finite-sample truncated quadratic risk difference[2] between the averaging estimator and the semiparametric estimator over a large class of DGPs. Extending the subsequence argument developed in Cheng, Liao, and Shi (2019) for generalized method of moments (GMM) estimators, I show that the sufficient conditions for the lower bound to be strictly less than zero and the upper bound to be weakly less than zero is mild. Since the class I consider includes DGPs under which the parametric restrictions are correctly specified, mildly misspecified and severely misspecified, my uniform dominance result asserts that the averaging estimator achieves improvement in accuracy over the semiparametric estimator in a way that is robust against misspecification. Unlike Cheng, Liao, and Shi (2019) who focus on one step GMM estimators, I consider two step M estimation framework for semiparametric models as it encompasses maximum likelihood estimator (MLE), GMM, many kernel-based and sieve estimators, etc. as special cases, as well as regular one step M estimators.

To demonstrate my averaging estimator, I present two carefully curated examples – a sample selection example and a quantile treatment effects (QTEs) example. I will introduce and then revisit both examples multiple times throughout this paper, demonstrating different aspects of my averaging estimator. A point worth emphasizing here is that even when the *estimation error* of the nonparametric component does not affect the asymptotic properties of the parametric component estimator (e.g., partially linear models in Robinson, 1988), the *presence* of the former and how it is *modeled* generally inflict critical impacts on the latter. This point will become clearer when I compare the two examples below.

This paper has a few 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 local 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. A two step procedure proposed by Claeskens and Hjort (2008) can be used to construct the confidence interval (also see, e.g., Kitagawa and Muris, 2016, for its application), but its coverage probability can be conservative. Third, I focus on averaging between one semiparametric estimator and one parametric estimator, excluding estimators that average the semiparametric estimator with more than one parametric estimators and potentially outperform the one proposed in this paper. These limitations all point out important directions for future research.

[2]The loss function and the truncated loss function are defined in equations (3.7) and (4.1), respectively.

Related Literature. This paper belongs to the growing literature on frequentist shrinkage and model averaging estimators, which are weighted averages of other estimators.[3] Shrinkage estimators date back to the James-Stein estimator in Gaussian models (James and Stein, 1961; Baranchik, 1964), and are comprehensively reviewed by Fourdrinier, Strawderman, and Wells (2018). Recent years have seen development of frequentist model averaging estimators in many contexts. Hjort and Claeskens (2003) and Hansen (2016) consider likelihood-based estimators in parametric models. In least square regression models, various model averaging estimator are developed and their properties are carefully examined by Judge and Mittelhammer (2004); Mittelhammer and Judge (2005); Hansen (2007); Wan, Zhang, and Zou (2010); Hansen and Racine (2012); Hansen (2014); Liu (2015) and Hansen (2017), just to name a few. Lu and Su (2015) study quantile regression models. For semiparametric models, Judge and Mittelhammer (2007); DiTraglia (2016) consider averaging GMM estimators, and Kitagawa and Muris (2016) analyze averaging semiparametric estimators of the treatment effects (ATT) based on different parametric propensity score models. Averaging estimators in nonparametric models are also considered, for example, by Fan and Ullah (1999); Yang (2001) and Yang (2003) and surveyed by Wasserman (2006, Chapter 7). Magnus, Powell, and Prüfer (2010) and Fessler and Kasy (2019), among others, investigate Bayesian model averaging estimators as well. Claeskens and Hjort (2008) provide an excellent review of both frequentist and Bayesian model averaging estimators. My paper differs from this literature in the following ways. First, compared with the papers on parametric and semiparametric models, I utilize a two step M estimation framework that nests many familiar estimators (one step or two step) as special cases. Second, in contrast to the literature on nonparametric models that deals with unknown functions and estimators with various rates of convergence, my paper focuses on finite dimensional parameters and a knife-edge case in which candidate estimators have the same rate of convergence. Third, my averaging weight, when specialized to corresponding cases, differs from those in the aforementioned papers. Fourth, I prove that my averaging estimator dominates the semiparametric benchmark using a uniform asymptotic approach, instead of the local asymptotic approach (Le Cam, 1972; Van der Vaart, 2000, Chapter 7) often taken in the literature. Finally, the sufficient condition for the uniform dominance of my averaging estimator, when certain weight matrix is chosen in the loss function (detailed in Section 4), is stronger than some estimators in the literature and weaker than others (detailed in Section 4).

This paper is particularly related to Cheng, Liao, and Shi (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

[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.

models.[4] Moreover, the restricted estimator considered in Cheng, Liao, and Shi (2019) is asymptotically efficient, but I allow the restricted (parametric) estimator to be away from the semiparametric efficiency bound. This relaxation is useful in practice since in complex semiparametric models, the efficient estimators under the parametric restrictions may be difficult to implement or may have certain undesirable features, and the widely used ones may fall short of the efficiency bound (such as Example 1).

The uniform asymptotic analysis in this paper premises upon high-level asymptotic distributions of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$, which can be justified under various low-level 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); Bickel and Ritov (2003); Powell (2001); Chen, Linton, and Van Keilegom (2003); Hirano, Imbens, and Ridder (2003); Firpo (2007); Newey (2009); Ichimura and Lee (2010); Ackerberg, Chen, and Hahn (2012); Ackerberg, Chen, Hahn, and Liao (2014) and Ichimura and Newey (2017) – and it is just impossible to enumerate all of them here.

Averaging estimators can be regarded as a smoothed generalization of pre-test estimators (or model selection estimators), as the latter restrict the averaging weights to be either zero or one depending on the result of certain specification test or criterion. For models involving infinite dimensional components, many authors propose various specification tests, including Bierens (1990), Wooldridge (1992), Hong and White (1995), Bierens and Ploberger (1997), Stinchcombe and White (1998), Li, Hsiao, and Zinn (2003) and Hart (2013) using sieve estimators, and Robinson (1989), Fan and Li (1996), Chen and Fan (1999), Lavergne and Vuong (2000), Aït-Sahalia, Bickel, and Stoker (2001), Horowitz and Spokoiny (2001), Fan, Zhang, and Zhang (2001) and Fan and Linton (2003) using kernel estimators. FIC-based model selection estimators in semiparametric models are considered in Hjort and Claeskens (2006); Claeskens and Carroll (2007); Zhang and Liang (2011); Vansteelandt, Bekaert, and Claeskens (2012) and DiTraglia (2016). Pre-test estimators typically perform better than the unrestricted benchmark for certain degrees of misspecification of the restrictions and worse for the others. Moreover, the literature has documented that in many settings, the maximal scaled quadratic risks of pre-test estimators based on consistent tests grow unbounded as sample sizes increase, despite promising properties suggested by pointwise asymptotic analysis. A well-cited example is the Hodges' estimator (e.g., Van der Vaart, 2000, Example 8.1), among others (Yang, 2005; Leeb and Pötscher, 2005, 2008; Hansen, 2016; Cheng, Liao, and Shi, 2019, etc.). In contrast, the uniform asymptotic approach of this paper better

[4]Cheng, Liao, and Shi (2019) is in turn based on the uniform inference analysis in Andrews, Cheng, and Guggenberger (2011).

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. Another direction is to provide valid inference for pre-test estimators (e.g., Belloni, Chernozhukov, and Hansen, 2014), but here I focus on developing estimator with uniform proved risks.

My 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; Robins and Rotnitzky, 2001; 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 is correctly specified, while my averaging estimator exhibits improved risk regardless of the degree of misspecification. Second, recent development in locally robust estimators in semiparametric models (e.g., Chernozhukov, Escanciano, Ichimura, Newey, and Robins, 2018; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018) removes impacts of the nuisance function estimation bias (brought by regularization of machine learning methods) on the influence function of the parameter of interest by orthogonalization (Neyman, 1959). My approach is still useful in light of their approach since how the nuisance function is *modeled* affects the influence function (both variance and bias) even when it is *known* and needs not estimation. Third, among the literature on sensitivity analysis (e.g., Rosenbaum and Rubin, 1983a; Leamer, 1985; Imbens, 2003; Altonji, Elder, and Taber, 2005; Andrews, Gentzkow, and Shapiro, 2017; Mukhin, 2018; Oster, 2019), Bonhomme and Weidner (2018) and Armstrong and Kolesár (2021) are the closest to my 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. My 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. In addition, I focus on point estimation while both the locally robust and the sensitivity analysis literature studies inference as well.

Plan of the Paper. The rest of this paper is organized as follows. Section 2 introduces the two examples. Section 3 describes my analysis framework, proposes my averaging weight and demonstrates using the two examples introduced in Section 2. Section 4 states and proves the main uniform dominance result of the paper, along with its conditions. Section 5 uses Monte Carlo simulations to investigate the finite sample performance of my averaging estimator in the two examples introduced in Section 2. Section 6 concludes. Appendix A provides more details on the two examples. Appendix B gives the proofs.

2 Examples

I use the following two examples to illustrate the implementation and the properties of my averaging estimator throughout this paper.

Example 1 - Sample Selection Model. *One is interested in estimating β in a partially linear model*

$$Y_{1i} = X_{1i}'\beta + s(X_{2i}) + V_{1i}, \tag{2.1}$$

where $\mathbb{E}(V_{1i}|X_{1i}, X_{2i}, Y_{2i} = 1) = 0$, X_{1i} is a $k \times 1$ vector, X_{2i} is an $l \times 1$ vector, Y_{2i} is an indicator variable, and X_{1i} and X_{2i} are assumed not to overlap for simplicity. Partially linear model may arise as a "reduced form" of the common sample selection model, where $s(X_{2i})$ is the correction term for the sample selection bias. For example,

$$Y_{1i} = \begin{cases} X_{1i}'\beta + U_{1i}, & \text{if } Y_{2i} = 1, \\ \text{unobserved}, & \text{if } Y_{2i} = 0, \end{cases} \tag{2.2}$$

$$Y_{2i} = \mathbb{I}\{X_{2i}'\gamma + U_{2i} \geq 0\}, \text{ where } \mathbb{I}\{\cdot\} \text{ is an indicator function.} \tag{2.3}$$

Equation (2.2) is often referred to as outcome equation, and (2.3) as selection equation.

If β in equation (2.1) is identified (see discussion on pages 5-8 of Ahn and Powell, 1993, and reference therein), then the estimator proposed by Robinson (1988) is one example of the semiparametric estimator $\hat{\beta}_{n,SP}$.[5] *If one imposes the parametric restriction that (U_{1i}, U_{2i}) are randomly drawn from a joint normal distribution, then Heckman (1979) shows that $s(X_{2i})$ is proportional to the "inverse Mill's ratio", and the widely used two step estimator suggested in his seminal paper could serves as one example of the parametric estimator $\hat{\beta}_{n,P}$.*[6]

Example 2 - Quantile Treatment Effects. *One is interested in estimating the QTEs at k different quantiles. Following the convention in the treatment effect literature, let T_i denote the treatment indicator, $Y_{1,i}$ and $Y_{0,i}$ denote unit i's potential outcomes with and without treatment. Let $q_{j,\tau} \equiv \inf\{q : P(Y_{j,i} \leq q) \geq \tau\}$ be the population τth quantile of the potential outcome $Y_{j,i}$ ($j = 0, 1$), and let $\Delta_\tau \equiv q_{1,\tau} - q_{0,\tau}$ denote the QTE at τ. In this example, $\beta \equiv (\Delta_{\tau_1}, \ldots, \Delta_{\tau_k})'$, where $\tau_\iota \in (0, 1)$ for $\iota = 1, \ldots, k$. Firpo (2007) shows the identification of $q_{j,\tau}$ under the standard strong ignorability conditions (Rosenbaum and Rubin, 1983b). In*

[5] Many semiparametric estimators of β in sample selection models (with potentially nonparametric selection equation) have been proposed and examined under various identification conditions, for example, Lee (1982); Gallant and Nychka (1987); Newey, Powell, and Walker (1990); Ahn and Powell (1993); Newey (2009), among many others. They could all serve as the $\hat{\beta}_{n,SP}$ in this paper, provided that Conditions in Section 4 are satisfied.

[6] Maximum likelihood estimator could also be applied under the joint normality restriction, and relative advantages of MLE and Heckman (1979) two step estimators are well studied (e.g., Wales and Woodland, 1980; Nelson, 1984).

terms of estimation, $\hat{q}_{j,\tau}$ solves the minimization problem (given j and τ):

$$\hat{q}_{n,j,\tau} \equiv \arg\min_{q} n^{-1} \sum_{i=1}^{n} \hat{\omega}_{n,j,i} \cdot \rho(Y_i - q), \tag{2.4}$$

where $\rho_\tau(a) \equiv a \cdot (\tau - \mathbb{I}\{a \leq 0\})$ is the check function often used for quantile regressions (e.g., Koenker and Bassett Jr, 1978), and the weights are functions of the estimated propensity score $\hat{p}_n(x)$ as follows

$$\hat{\omega}_{n,1,i} \equiv \frac{T_i}{\hat{p}_n(X_i)}, \text{ and } \hat{\omega}_{n,0,i} \equiv \frac{1 - T_i}{1 - \hat{p}_n(X_i)}. \tag{2.5}$$

Depending on whether the propensity score function $p(x)$ is estimated parametrically (e.g., probit) or nonparametrically (e.g., series logit in Firpo, 2007), one would obtain a parametric estimator $\hat{\beta}_{n,P}$ or a semiparametric estimator $\hat{\beta}_{n,SP}$ of the QTEs.

The semiparametric models considered in this paper is flexible enough to include many other examples – control function approach to endogenous regressors (Blundell and Powell, 2004), single-index models (Ahn, Ichimura, and Powell, 1996), dynamic games (Bajari, Chernozhukov, Hong, and Nekipelov, 2015) and dynamic discrete choice models (Hotz and Miller, 1993; Buchholz, Shum, and Xu, 2020; Keane and Wolpin, 1997), etc.

I put Examples 1 and 2 in the spotlight here because they highlight a few distinct features of my averaging estimator. First, unlike in Cheng, Liao, and Shi (2019), the restricted estimator $\hat{\beta}_{n,P}$ in this paper need not to be asymptotically efficient under the parametric restrictions, which is illustrated by Heckman (1979) two step estimator in Example 1.[7] 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) – this is the case in both Examples 1 and 2 – even though it might be invariant to the *estimation error* of the first step nuisance parameter like in Example 1. Third, in the particular Monte Carlo experiments in Section 5, the relative efficiency gain of $\hat{\beta}_{n,P}$ compared to $\hat{\beta}_{n,SP}$ outweighs its misspecification bias in Example 1, where the averaging weight is strictly between 0 and 1 and the averaging estimator $\hat{\beta}_{n,\hat{w}}$ strictly dominates $\hat{\beta}_{n,SP}$; opposite is the case in Example 2, where the averaging weight is close to 0 and the averaging estimator exhibit the same asymptotic properties as $\hat{\beta}_{n,SP}$ (i.e., only weakly dominates $\hat{\beta}_{n,SP}$).

[7] In contrast, the less frequently used MLE is an asymptotically efficient parametric estimator. Further discussion on the two estimators is in Appendix A.

3 Framework of Analysis and Averaging Weight

In this section, I describe the general framework of my analysis, prescribe the averaging weight, and explain its intuition. How to obtain the averaging weight in a particular semiparametric model is then demonstrated using Examples 1 and 2.

General Framework. I am 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 $\mathcal{F}$. Suppose β_F, the true parameter value under DGP F, is identified as the unique minimizer of some objective function $Q_F(\beta, h_F)$; that 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 argument, the presence of h and how it is modeled generally affect the asymptotic properties of β estimators through Q_F, even in the absence of estimation error of h.[8] Under DGP F, the true nuisance parameter value h_F is identified as the unique maximizer of another objective function $R_F(h)$; that is,

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

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

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

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

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

Depending on how h is estimated in the first step, I may end up with different estimators of β. If I do not impose specific functional form restrictions on h, then $\hat{h}_n$ can be obtained using common nonparametric estimation procedures. For instance, $\hat{h}_n$ may result from a first step sieve M estimation procedure as follows,

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

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

where $\hat{R}_n(h)$ is some objective function, and $\mathcal{H}_n$ are subspaces of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ that become dense as $n \to \infty$. The semiparametric estimator $\hat{\beta}_{n,SP}$ thus results from a two step M estimation procedure with the first step being (3.4) and the second step being (3.3).

On the other hand, economic hypotheses may suggest certain parametric form of h, or one may want to limit the dimension of h to improve the efficiency. Whatever the motive might be, one can model h with a finite dimensional subspace of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, denoted as $\mathcal{H}_g$, with a function g that is known up to a finite dimensional vector of unknown parameters γ. Formally, let $\Gamma \subset \mathbb{R}^t$ be a compact subset of the t-dimensional Euclidean space, and let $\gamma \in \Gamma$, then

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

Let

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

and let the restricted nuisance parameter estimate be written as $\hat{h}_n \equiv g_{\hat{\gamma}_n}$, then the parametric estimator $\hat{\beta}_{n,P}$ results from a two step M estimation procedure with the first step being (3.6) and the second step being (3.3).

Heuristics and Averaging Weight. For any estimator $\hat{\beta}_n$ of β, I consider a quadratic loss function.[9] For a chosen positive semi-definite weight matrix Υ, I define the loss function to be

$$\ell(\hat{\beta}_n, \beta) \equiv n(\hat{\beta}_n - \beta)'\Upsilon(\hat{\beta}_n - \beta). \tag{3.7}$$

Here the weight matrix Υ is chosen by the researcher and reflects how much the researcher values the estimation accuracy of each coordinate of β. If the researcher treats every coordinate equally, then she may choose $\Upsilon = I_k$ (the $k\times k$ identity matrix). If the researcher focuses on the prediction error in the sample selection model, then she may choose $\Upsilon = \mathbb{E}_F(X_{1i}X_{1i}')$, where $\mathbb{E}_F(\cdot)$ denotes the expectation operator under DGP F. If the researcher focuses on only a subvector of β, then she may choose Υ to be a diagonal matrix with diagonal entries associated with the subvector being one and other diagonal entries being zero. This last example shares the same spirit with the focused information criterion (FIC) model averaging (Zhang and Liang, 2011), but the weight matrix Υ here affords more flexibility. Note that both the loss function and the averaging weight (to be introduced later) depend on Υ, but I suppress such dependence for notational simplicity.

Given the loss function in equation (3.7), the semiparametric estimator $\hat{\beta}_{n,SP}$ is preferred

[9] 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 to be defined in equation (4.1).

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 be more efficient than $\hat{\beta}_{n,SP}$ since the parametric first step $g_{\hat{\gamma}_n}$ are generally more efficient than the nonparametric first step $\hat{h}_n$. As a result, the potentially more efficient $\hat{\beta}_{n,P}$ sometimes has improved 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 measure unknown to the researcher.

The main message of this paper, therefore, is that with the averaging weight I propose, the averaging estimator of the form in equation (1.1) *always* has no larger risk than the robust estimator $\hat{\beta}_{n,SP}$ *regardless* of whether the parametric restrictions hold or not. I prescribe the averaging weight and explain the heuristics in this section, and rigorous conditions and the formal uniform dominance result will be provided in Section 4.

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 cov_F be their asymptotic covariance matrix. Let $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ and $\widehat{cov}_n$ be their consistent estimators. Then the data-driven averaging weight is

$$\hat{w}_n \equiv \frac{\text{tr}[\Upsilon(\widehat{V}_{n,SP} - \widehat{cov}_n)]}{\text{tr}[\Upsilon(\widehat{V}_{n,SP} + \widehat{V}_{n,P} - 2\widehat{cov}_n)] + n(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})'\Upsilon(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})}, \tag{3.8}$$

where $\text{tr}(\cdot)$ indicates the trace of a square matrix.

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

$$\hat{w}_n \equiv \frac{\text{tr}[\Upsilon(\widehat{V}_{n,SP} - \widehat{V}_{n,P})]}{\text{tr}[\Upsilon(\widehat{V}_{n,SP} - \widehat{V}_{n,P})] + n(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})'\Upsilon(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})}, \tag{3.9}$$

which resembles the GMM averaging weight proposed by Cheng, Liao, and Shi (2019). It is easier to see the intuition of the averaging weight from equation (3.9). If the efficiency gain of imposing the first step parametric restrictions, represented by $\text{tr}[\Upsilon(\widehat{V}_{F,SP} - \widehat{V}_{F,P})]$, is large, then the averaging estimator ought to allocate more weight to $\hat{\beta}_{n,P}$. If, on the other hand, the 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 (3.9) operationalizes such intuition by striking a balance between robustness and efficiency in a GMM setting.

The weight in equation (3.8) generalizes (3.9) by allowing 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 Heckman (1979) two step estimator is asymptotically inefficient but more widely used than the efficient MLE, for a variety of reasons (see, e.g., the discussion in Heckman, 1976; Wales and Woodland, 1980; Nelson, 1984).

In the rest of this section, I will use Examples 1 and 2 to demonstrate the construction of the averaging weight $\hat{w}_n$. Here I will assume that $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are both asymptotically linear estimators (equation (2.1) of Ichimura and Newey, 2017). Asymptotic linearity is not needed for the main dominance result in this paper,[10] but if it holds, then the consistent estimators $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ and $\widehat{cov}_n$ in equation (3.8) can be readily obtained based on the influence functions of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$. Let $\psi_{F,SP}(z)$ denote the non-centered influence function of $\hat{\beta}_{n,SP}$, let $\psi_{F,P}(z)$ denote that of $\hat{\beta}_{n,P}$, and let $\psi_{n,SP}$ and $\psi_{n,P}$ denote their sample analogs, respectively.[11] Then

$$\hat{V}_{n,SP} = \frac{1}{n}\sum_{i=1}^{n} \psi_{n,SP}(Z_i)\psi_{n,SP}'(Z_i) - \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,SP}(Z_i)\right] \cdot \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,SP}(Z_i)\right]', \tag{3.10}$$

$$\hat{V}_{n,P} = \frac{1}{n}\sum_{i=1}^{n} \psi_{n,P}(Z_i)\psi_{n,P}'(Z_i) - \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,P}(Z_i)\right] \cdot \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,P}(Z_i)\right]', \tag{3.11}$$

$$\widehat{cov}_n = \frac{1}{n}\sum_{i=1}^{n} \psi_{n,SP}(Z_i)\psi_{n,P}'(Z_i) - \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,SP}(Z_i)\right] \cdot \left[\frac{1}{n}\sum_{i=1}^{n} \psi_{n,P}(Z_i)\right]'. \tag{3.12}$$

It is worth emphasizing that the influence functions need to be valid under *potential misspecification* (e.g., Ichimura and Lee, 2010), such that the estimators $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ and $\widehat{cov}_n$ are consistent *regardless* of whether the parametric restrictions hold or not. In other words, they have to be robust against misspecification of the parametric restrictions; otherwise the resulting averaging estimator might not uniformly dominate the semiparametric benchmark. In particular, the influence functions $\psi_{F,SP}(z)$ and $\psi_{F,P}(z)$ depend on the unknown DGP F, and such dependence is often manifested in the fact that $\psi_{F,SP}(z)$ and $\psi_{F,P}(z)$ involve the unknown parameter β_F itself (or other functionals of the DGP F). Whenever β_F appears in the influence functions, the robust estimator $\hat{\beta}_{n,SP}$ (or robust estimators of the functionals) needs to be used.

Example 1 (cont'd) - Sample Selection Model. *Let $h_F = (h_{1F}, h_{2F}')'$ be a function from $\mathbb{R}^l$ to $\mathbb{R}^{k+1}$, with $h_{1F}(x_2) \equiv \mathbb{E}_F(Y_1|X_2 = x_2)$ and $h_{2F}(x_2) \equiv \mathbb{E}_F(X_1|X_2 = x_2)$. Let $\hat{h}_n$ denote*

[10] Nearly all root-n consistent semiparametric estimators are asymptotically linear under sufficient regularity conditions (Bickel, Klaassen, Bickel, Ritov, Klaassen, Wellner, and Ritov, 1993; Ichimura and Newey, 2017). See the discussion after Condition 2 for more details.

[11] This facilitates the demonstration by giving the readers some concrete objects to look at.

the Nadaraya-Watson estimator of h_F. Let $\hat{\lambda}_i \equiv \lambda(-X_{2i}'\hat{\gamma}_n)$ be the estimated inverse Mill's ratio with $\hat{\gamma}_n$ being the first step probit estimate associated with $\hat{\beta}_{n,P}$, and let $P_{\hat{\lambda}}(Y_{1i})$ and $P_{\hat{\lambda}}(X_{1i})$ denote the fitted values after regressing Y_{1i} and X_{1i} respectively on $\hat{\lambda}_i$. Since h does not depend on β, I use Theorem 3.3 of Ichimura and Lee (2010) to derive the non-centered influence functions of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ under potential misspecification (detailed in Appendix A) and immediately get their sample analogs:

$$\psi_{n,SP}(Z_i) = -\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}\left[X_{1i} - \hat{h}_{2,n}(X_{2i})\right]\left[X_{1i} - \hat{h}_{2,n}(X_{2i})\right]'\right\}^{-1}$$
$$\cdot \{Y_{1i} - \hat{h}_{1,n}(X_{2i}) - [X_{1i} - \hat{h}_{2,n}(X_{2i})]'\hat{\beta}_{n,SP}\} \cdot [X_{1i} - \hat{h}_{2,n}(X_{2i})], \quad (3.13)$$

$$\psi_{n,P}(Z_i) = -\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}\left[X_{1i} - P_{\hat{\lambda}}(X_{1i})\right]\left[X_{1i} - P_{\hat{\lambda}}(X_{1i})\right]'\right\}^{-1}$$
$$\cdot \{Y_{1i} - P_{\hat{\lambda}}(Y_{1i}) - [X_{1i} - P_{\hat{\lambda}}(X_{1i})]'\hat{\beta}_{n,SP}\} \cdot (X_{1i} - P_{\hat{\lambda}}(X_{1i})), \quad (3.14)$$

where n_1 is the size of the subsample whose Y_{1i} is observed and n is the total sample size. As a result, the averaging weight can be constructed by first plugging equations (3.13) *and* (3.14) *into equations (3.10), (3.11) and* (3.12)*, and then plugging the latter into equation* (3.8)*, with the coefficient n of the second term of the denominator in equation* (3.8) *replaced by n_1.*

Two point are worth emphasizing here. First, β naturally arises in the non-centered influence functions (see details in Appendix A) and is invariant to how the conditional mean function h is modeled. As a result, when computing the sample analogs of the non-centered influence functions using equations (3.13) and (3.14), β should be replaced by $\hat{\beta}_{n,SP}$, the estimator that is consistent regardless of whether the joint normality restriction is correctly specified or not. Second, the nuisance function h directly enters the influence function of $\hat{\beta}_n$. As a result, how h is modeled (by nonparametric $h_F(x_2)$ or by projection P_{λ_F} on the inverse Mill's ratio) affects the functional form of the influence function, even though neither equation (3.13) nor equation (3.14) contains a correction term for the first step estimation error of h.

Example 2 (cont'd) - (Quantile Treatment Effects). *Let $\hat{p}_{n,SP}(x)$ denote the series logit estimator of the propensity score function $p(x)$ proposed by Firpo (2007), and let $\Phi(x'\hat{\gamma}_n)$ denote the linear probit estimator of $p(x)$, with $\hat{\gamma}_n$ being the probit coefficient estimate. Note that $\hat{\beta}_{n,SP} = \hat{q}_{1,n,SP} - \hat{q}_{0,n,SP}$ and $\hat{\beta}_{n,P} = \hat{q}_{1,n,P} - \hat{q}_{0,n,P}$, where $\hat{q}_{j,n,SP} = (\hat{q}_{j,\tau_1,n,SP}, \ldots, \hat{q}_{j,\tau_k,n,SP})'$ and $\hat{q}_{j,P} = (\hat{q}_{j,\tau_1,n,P}, \ldots, \hat{q}_{j,\tau_k,n,P})'$. Here $\hat{q}_{j,\tau,n,SP}$ denotes the semiparametric estimator of the τth quantile of $Y_{j,i}$ ($j = 0, 1$), obtained by solving equation* (2.4) *with $\hat{p}_n(X_i)$ in equation* (2.5) *replaced by $\hat{p}_{n,SP}(X_i)$; and $\hat{q}_{j,\tau,n,P}$ denotes the corresponding parametric estimator by*

replacing $\hat{p}_n(X_i)$ *in equation* (2.5) *with* $\Phi(X_i'\hat{\gamma}_n)$.[12] *I follow the argument in the proof of Theorem 1 in Firpo (2007, Appendix 3) to derive the influence functions of* $\hat{\beta}_{n,SP}$ *and* $\hat{\beta}_{n,P}$ *(detailed in Appendix A) and immediately get their sample analogs for any* $\tau \in (0,1)$*:*

$$\begin{aligned}
\hat{\pi}_{j,\tau,n,SP}(y) &\equiv -\frac{\mathbb{I}\{y \leq \hat{q}_{j,\tau,n,SP}\} - \tau}{\hat{f}_{j,n,SP}(\hat{q}_{j,\tau,n,SP})}, j = 0,1, \\
\hat{\varphi}_{\tau,n,SP}(Z_i) &\equiv \frac{T_i}{\hat{p}_{n,SP}(X_i)} \cdot \hat{\pi}_{1,\tau,n,SP}(Y_i) - \frac{1-T_i}{1-\hat{p}_{n,SP}(X_i)} \cdot \hat{\pi}_{0,\tau,n,SP}(Y_i), \\
\hat{\alpha}_{\tau,n,SP}(Z_i) &\equiv -\hat{\mathbb{E}}_{n,SP}\left[\frac{T \cdot \hat{\pi}_{1,\tau,SP}(Y)}{[\hat{p}_{n,SP}(X)]^2} + \frac{(1-T)\cdot \hat{\pi}_{0,\tau,SP}(Y)}{[1-\hat{p}_{n,SP}(X)]^2}\middle| X = X_i\right] \cdot (T_i - \hat{p}_{n,SP}(X_i)), \\
\hat{\psi}_{\tau,n,SP}(Z_i) &\equiv \hat{\varphi}_{\tau,n,SP}(Z_i) + \hat{\alpha}_{\tau,n,SP}(Z_i);
\end{aligned} \tag{3.15}$$

and

$$\begin{aligned}
\hat{\pi}_{j,\tau,n,P}(y) &\equiv -\frac{\mathbb{I}\{y \leq \hat{q}_{j,\tau,n,P}\} - \tau}{\hat{f}_{j,n,SP}(\hat{q}_{j,\tau,n,P})}, j = 0,1, \\
\hat{\varphi}_{\tau,n,P}(Z_i) &\equiv \frac{T_i}{\Phi(X_i'\hat{\gamma}_n)} \cdot \hat{\pi}_{1,\tau,n,P}(Y_i) - \frac{1-T_i}{1-\Phi(X_i'\hat{\gamma}_n)} \cdot \hat{\pi}_{0,\tau,n,P}(Y_i), \\
\hat{\alpha}_{\tau,n,P}(Z_i) &\equiv -\hat{\mathbb{E}}\left[\frac{T_i \cdot \hat{\pi}_{1,\tau,n,P}(Y)}{[\Phi(X_i'\hat{\gamma}_n)]^2} + \frac{(1-T_i)\cdot \hat{\pi}_{0,\tau,n,P}(Y)}{[1-\Phi(X_i'\hat{\gamma}_n)]^2}\middle| X = X_i\right] \cdot (T_i - \Phi(X_i'\hat{\gamma}_n)), \\
\hat{\psi}_{\tau,n,P}(Z_i) &\equiv \hat{\varphi}_{\tau,n,P}(Z_i) + \hat{\alpha}_{\tau,n,P}(Z_i),
\end{aligned} \tag{3.16}$$

where $\hat{f}_{j,n,SP}(\cdot)$ *is a nonparametric estimator of the probablity density function of the potential outcome* Y_j *(*$j = 0,1$*) and* $\hat{\mathbb{E}}_{n,SP}(\cdot)$ *is a nonparametric estimator of the conditional mean function, both are described in details in Appendix 3 of Firpo (2007). Let* $\hat{\psi}_{n,SP}(Z_i) \equiv \left(\hat{\psi}_{\tau_1,n,SP}(Z_i), \ldots, \hat{\psi}_{\tau_k,n,SP}(Z_i)\right)'$ *and* $\hat{\psi}_{n,P}(Z_i) \equiv \left(\hat{\psi}_{\tau_1,n,P}(Z_i), \ldots, \hat{\psi}_{\tau_k,n,P}(Z_i)\right)'$*. As a result, the averaging weight can be constructed by first plugging* $\hat{\psi}_{n,SP}(Z_i)$ *and* $\hat{\psi}_{n,P}(Z_i)$ *into equations (3.10), (3.11) and* (3.12)*, and then plugging the latter into equation* (3.8)*.*

Similar to Example 1, it is worth emphasizing that in both sample analogs, $\hat{f}_{j,n,SP}(\cdot)$ and $\hat{\mathbb{E}}(\cdot)$ are both nonparametric estimators, since $f_j(\cdot)$ and $\mathbb{E}(\cdot)$ naturally arise in the influence functions from approximating the optimization problem in equation (2.4) by a quadratic optimization problem with coefficients involving the true probability density function under DGP F (equation (A.1) and (A.2) in Appendix A), and therefore are invariant to how the propensity score function $p(\cdot)$ is modeled.

In contrast to Example 1, both influence functions in equations (3.15) and (3.16) contain a correction term (i.e., $\hat{\alpha}_{\tau,n,SP}(Z_i)$ and $\hat{\alpha}_{\tau,n,P}(Z_i)$) for the first step estimation error of h, in addition to the terms (i.e., $\hat{\varphi}_{\tau,n,SP}(Z_i)$ and $\hat{\varphi}_{\tau,n,P}(Z_i)$) as if the propensity score function was

[12]The details of the procedure for obtaining $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are elaborated in Appendix A.

known. The functional form of all of them depends on how the propensity score function is modeled (by nonparametric $p_F(x)$ or by $\Phi(x'\gamma_F)$, the probit probability).

4 Main Result

In this section, I prove and provide the conditions for the uniform dominance result of the averaging estimator. That is, in the two step M estimation framework, the averaging estimator $\hat{\beta}_{n,\hat{w}}$ proposed in equation (1.1) with the weight given in equation (3.8) has (weakly) smaller asymptotic quadratic risk than the robust semiparametric estimator $\hat{\beta}_{n,SP}$ under DGPs in $\mathcal{F}$, which encompasses a wide range of DGPs under which the parametric restrictions might be correctly specified or misspecified.

The key 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. I utilize the uniform asymptotic approach and the subsequence technique in Cheng, Liao, and Shi (2019), instead of Pitman sequences, which is frequently used when analyzing the local asymptotic properties of estimators. Lower (infimum) and upper (supremum) bounds of the risk differences between $\hat{\beta}_{n,\hat{w}}$ and $\hat{\beta}_{n,SP}$ for all DGPs within a set $\mathcal{F}$ satisfying certain regularity conditions are considered before rendering the sample size to infinity.

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

$$\ell_\zeta(\hat{\beta}_n, \beta) \equiv \min\{\ell(\hat{\beta}_n, \beta), \zeta\}, \tag{4.1}$$

where $\ell(\hat{\beta}_n, \beta)$ is the quadratic loss function defined in equation (3.7). This truncated loss function is defined to facilitate my asymptotic analysis later, and the truncation does not restrict the applicability of the main result much since ζ could be arbitrarily large. The bounds of the truncated risk differences for finite sample size n are defined as:

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

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

$$Asy\underline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \equiv \lim_{\zeta\to\infty} \liminf_{n\to\infty} \underline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta), \tag{4.2}$$

$$Asy\overline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \equiv \lim_{\zeta\to\infty} \limsup_{n\to\infty} \overline{RD}_n(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}; \zeta). \tag{4.3}$$

Note that first the extrema of the risk differences over the entire DGP set $\mathcal{F}$ are taken, then the sample size is sent to infinity. The averaging estimator is said to dominate the semiparametric estimator in terms of asymptotic truncated risk uniformly over $\mathcal{F}$ if

$$Asy\underline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) < 0, \tag{4.4}$$

and

$$Asy\overline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \leq 0. \tag{4.5}$$

For every DGP $F \in \mathcal{F}$ and under the parametric restrictions, define the first step pseudo-true parameter vector as

$$\gamma_F \equiv \arg\min_{\gamma \in \Gamma} R_F(g_\gamma), \tag{4.6}$$

where the first step objective function $R_F(\cdot)$ is the same as in equation (3.2) and the first step nuisance function subspace $\mathcal{H}_g$ is defined in equation (3.5). Also define the second step pseudo-true parameter as

$$\beta_{F,P} \equiv \arg\min_{\beta \in \mathcal{B}} Q_F(\beta, g_{\gamma_F}), \tag{4.7}$$

where $Q_F(\cdot,\cdot)$ is the same as in equation (3.1). In general, the nuisance function g_{γ_F} induced by the pseudo-true parameter γ_F is different from the true nuisance function h_F identified in (3.2). In consequence, $\beta_{F,P}$ in general will be different from β_F, the true parameter of interest identified in (3.1).

Define $\delta_F \equiv \beta_{F,P} - \beta_F$, which represents the bias caused by imposing the parametric restrictions.

Condition 1. Suppose $\mathcal{F}$ is such that the following holds.
(i) $\inf_{\{F \in \mathcal{F}:\ \|g_{\gamma_F} - h_F\|_{\mathcal{H}} > 0\}} \frac{\|\delta_F\|}{\|g_{\gamma_F} - h_F\|_{\mathcal{H}}} > 0$;
(ii) $0_{k\times 1} \in \text{int}(\Delta_\delta)$, 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 is misspecified, then the pseudo-true parameter value $\beta_{F,P}$ will differ from the true value β_F, which rules out the uninteresting special case that β_F may be consistently estimable even with severely misspecified parametric restrictions. As a result, the degree of misspecification can be indexed by δ_F, the bias introduced by imposing the parametric restrictions. Condition 1(ii) says that the parametric restrictions may be misspecified of varied degrees, including the correct specification case. Condition 1 does not impose any stringent restrictions on the models that I analyze.

Recall that $V_{F,SP}$ and $V_{F,P}$ are the asymptotic variance-covariance matrices of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ under DGP F, respectively, and cov_F is their asymptotic covariance matrix. I use $S(F)$ to denote the nuisance parameter vector that characterizes the joint asymptotic distributions

of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ under DGP F; that is,

$$S(F) \equiv [\delta_F', \text{vech}(V_{F,SP})', \text{vech}(V_{F,P})', \text{vec}(cov_F)']'.$$

Let $\bar{S}(F)$ denote the subvector of $S(F)$ excluding δ_F and define

$$\mathcal{S} \equiv \{S(F):\ F \in \mathcal{F}\}. \tag{4.8}$$

Condition 2. For a sequence of DGPs $\{F_n\}_{n=1}^{\infty}$ such that $S(F_n) \to S(F)$ 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.[13]

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

$$\begin{bmatrix} n^{1/2}(\hat{\beta}_{n,SP} - \beta_{F_n}) \\ n^{1/2}(\hat{\beta}_{n,P} - \beta_{F_n}) \end{bmatrix} \xrightarrow{d.} \begin{bmatrix} \xi_{F,SP} \\ \xi_{F,P} + d \end{bmatrix}. \tag{4.9}$$

If I define $\tilde{\xi}_F \equiv (\xi_{F,SP}', \xi_{F,P}')'$ and

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

then $\tilde{\xi}_F \sim \mathcal{N}(0_{2k\times 1}, \tilde{V}_F)$, with $V_{F,SP} \geq V_{F,P}$.

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

Condition 2(i) requires that both $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are locally regular estimators (Ichimura and Newey, 2017, Definition 1), which means that $n^{1/2}((\hat{\beta}_{n,SP}', \hat{\beta}_{n,P}')' - (\beta_{F_n}', \beta_{F_n,P}')')$ has the same limiting distribution under a sequence of alternatives as it does when $F_n = F$ for all n. Like in Ichimura and Newey (2017, Section 3), this condition is a mild one and it allows me to bypass primitive conditions of asymptotic linearity (e.g., Ritov and Bickel, 1990) and to focus on the main dominance result of this paper. Note that in equation (4.9), $\hat{\beta}_{n,P}$ is re-centered using δ_{F_n} and the presumption that $n^{1/2}\delta_{F_n} \to d$. Moreover, $V_{F,SP} \geq V_{F,P}$ states the intuition that imposing parametric restrictions generally lead to (weak) efficiency gain.[14] This intuitive condition can be justified by the Le Cam's Third Lemma (e.g., Van der Vaart, 2000, Example 6.7) and the definition of semiparametric information bound (see Bickel, Klaassen, Bickel, Ritov, Klaassen, Wellner, and Ritov, 1993, Chapter 3) as follows. First, when $\|d\| = 0$, the parametric restrictions are correctly specified, due to Condition 1(ii). As a result, 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

[13] Note that Pitman sequences are an example of such sequences.

[14] For the two examples, this can also be seen from the influence functions given in Appendix A.

Ackerberg, Chen, Hahn, and Liao (2014), one can show that the semiparametric efficiency bound of the restricted model (with nuisance function space $\mathcal{H}_g$) is smaller than that of the unrestricted model (with nuisance function space $\mathcal{H}$),[15] because the latter is the supremum of all parametric submodels that include the former. So it is natural to require that $V_{F,SP} \geq V_{F,P}$.[16] Second, when $\|d\| < \infty$ but $\|d\| \neq 0$, the asymptotic variance-covariance matrix of $\hat{\beta}_{n,P}$ remains $V_{F,P}$ by the local regularity and the 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, $V_{F,SP} \geq V_{F,P}$ follows. 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 in Appendix B. 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, Imbens, and Ridder, 2003, are pertinent to Examples 1 and 2).[17] I bypass those conditions and focus 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 equations (3.1) and (4.7) as presumption, for which the primitive conditions have been studied extensively (e.g., Newey and McFadden, 1994, Section 2).

Note that the nuisance parameter $S(F)$ determines the asymptotic properties of $\hat{w}_n$ and $\hat{\beta}_{n,\hat{w}_n}$ through those of $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$. Given the high-level Condition 2, the following lemma follows immediately.

Lemma 1. *Suppose Conditions 1 and 2 hold and let* $A_F \equiv \Upsilon(V_{F,SP} - cov_F)$ *and* $B_F \equiv \Upsilon(V_{F,SP} + V_{F,P} - 2cov_F)$. *Suppose that* $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ *and* $\widehat{cov}_n$ *have finite probability limits.*

(i) If $\|d\| < \infty$, *note that* $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ *and* $\widehat{cov}_n$ *are consistent estimators of corresponding*

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

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

[17] In Cheng, Liao, and Shi (2019), additional structure brought by the GMM facilitates discussion on low-level conditions.

elements in $\bar{S}(F)$*, then*

$$\hat{w}_n \xrightarrow{d.} w_F \equiv \frac{tr(A_F)}{tr(B_F) + (\xi_{F,P} + d - \xi_{F,SP})'\Upsilon(\xi_{F,P} + d - \xi_{F,SP})}, \tag{4.10}$$

which in turn implies that

$$n^{1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_F) \xrightarrow{d.} \bar{\xi}_{F,d} \equiv (1 - w_F)\xi_{F,SP} + w_F(\xi_{F,P} + d); \tag{4.11}$$

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

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

(i) $\mathcal{S}$ is compact;

(ii) for any $F \in \mathcal{F}$ with $\delta_F = 0$, 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 $C, \kappa > 0$.

Condition 3(i) is necessary for applying the subsequence argument to show the uniform dominance result. Since $\mathcal{S}$ is a set of finite dimensional vectors, Condition 3(i) essentially requires $\mathcal{S}$ to be bounded and closed. vech$(V_{F,SP})$ and vech$(V_{F,P})$ being bounded implies that both $\hat{\beta}_{n,SP}$ and $\hat{\beta}_{n,P}$ are regular estimators, which is satisfied in most cases I am interested in (e.g., Examples 1 and 2). $\mathcal{S}$ being closed is not restrictive in the sense that if $\mathcal{S}$ is not closed, then I can define it to be the closure of $\mathcal{S}$ and the main uniform dominance result still holds. Condition 3(ii) says that for any $F \in \mathcal{F}$ satisfying the parametric restrictions, there are many DGPs $\tilde{F} \in \mathcal{F}$ that are close to F, where the closeness of two DGPs is measured by the distance between $\bar{S}(\tilde{F})$ and $\bar{S}(F)$. 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.

Now I explain the rationale behind the averaging weight $\hat{w}_n$ in equation (3.8). 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:

$$n^{1/2}(\hat{\beta}_{n,w} - \beta_F) \xrightarrow{d.} \xi_{F,w} \equiv (1 - w)\xi_{F,SP} + w(\xi_{F,P} + d). \tag{4.12}$$

Since the asymptotic risk, defined in equation (3.7), is quadratic in w, the optimal weight w^* that minimizes the asymptotic risk under DGP F is

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

If $\hat{\beta}_{n,P}$ is asymptotically efficient under the parametric restrictions, then $cov_F = V_{F,P}$ and

the optimal weight simplifies to

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

This optimal weight balances the efficiency gain and the bias induced by the first step parametric restrictions. Higher efficiency gain $\text{tr}[\Upsilon(V_{F,SP} - V_{F,P})]$, relative to the squared bias $d'\Upsilon d$, demands larger weight w^* to be assigned to $\hat{\beta}_{n,P}$, and vice versa.

Although the optimal weight w^* is infeasible due to unknown $V_{F,SP}$, $V_{F,P}$, cov_F and d, equation (3.8) constructs a feasible averaging weight by replacing the unknown components with their estimators. In equation (3.8), $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ and $\widehat{cov}_n$ are consistent estimators of $V_{F,SP}$, $V_{F,P}$ and cov_F, respectively. At the same time, Condition 2(i) implies that $n^{1/2}(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP})$ is an asymptotically unbiased estimator of d when $\|d\| < \infty$, so $d'\Upsilon d$ in w^* is further replaced by $n\left(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP}\right)' \Upsilon \left(\hat{\beta}_{n,P} - \hat{\beta}_{n,SP}\right)$ in order to get equation (3.8).

When $\|d\| = \infty$, the parametric estimator $\hat{\beta}_{n,P}$ is severely biased so that a sensible averaging estimator ought to allocate zero weight to $\hat{\beta}_{n,P}$. This intuition is echoed by Condition 2(ii) and Lemma 1(ii), which imply that the feasible averaging weight given in equation (3.8) approaches to zero, as long as the probability limits of $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ and $\widehat{cov}_n$ are finite.

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 Lemma 1(i) is a random variable and in general not unbiased for w^* due to the Jensen's inequality, so $\hat{w}_n$ is not a consistent estimator for the infeasible optimal weight w^*. Proving the uniform dominance of the averaging estimator, therefore, is more challenging than it might appear at first sight, since $\hat{\beta}_{n,SP}$, $\hat{\beta}_{n,P}$ and $\hat{w}_n$ are mutually dependent random variables and their randomness needs to be dealt with at the same time. For this purpose, I will utilize the subsequence technique employed by Cheng, Liao, and Shi (2019).

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

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

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

$$\mathcal{U} \equiv \{u_{F,d}\text{: } \|d\| < \infty, \text{ and } F \in \mathcal{F} \text{ with } \delta_F = 0]\}. \tag{4.13}$$

and

$$\mathcal{U}_\infty \equiv \{u_{F,d}\text{: } \|d\| = \infty, \text{ and } F \in \mathcal{F}\}. \tag{4.14}$$

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

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

$$r(u_{F,d}) \equiv \mathbb{E}\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} - \xi'_{F,SP}\Upsilon\xi_{F,SP}\right),$$

where $\bar{\xi}_{F,d}$ and $\xi_{F,SP}$ are defined in equations (4.11) and (4.9), respectively. $\mathcal{U}$ and $\mathcal{U}_\infty$ defined here may appear similar to the set $\mathcal{S}$ defined in equation (4.8), but they are different. For any $u_{F,d} \in \mathcal{U} \cup \mathcal{U}_\infty$, the corresponding $\delta \equiv n^{-1/2}d$ is a different object from δ_F associated with F. $\mathcal{S}$ is the set of *actual* nuisance parameter vectors that determine the asymptotic properties of $\hat{\beta}_{n,SP}$, $\hat{\beta}_{n,P}$ and $\hat{\beta}_{n,\hat{w}_n}$ under DGPs in $\mathcal{F}$. In contrast, $\mathcal{U}$ is the set of all *hypothetical* nuisance parameter vectors that would have prevailed had the asymptotic variance-covariance matrices $V_{F,SP}$, $V_{F,P}$ and cov_F been the same as some DGP with zero bias ($\delta_F = 0$) from $\mathcal{F}$ and had the asymptotic bias d been finite. Note that if $u_{F,d} \in \mathcal{U}$ (i.e., $\|d\| < \infty$), the corresponding δ ranges from being zero to approaching to infinity at any rate that is not faster than $n^{1/2}$, corresponding to *correct* specification or *mild* misspecification of the parametric restrictions. Similarly, $\mathcal{U}_\infty$ is the set of all *hypothetical* nuisance parameter vectors that would have prevailed had the asymptotic variance-covariance matrices $V_{F,SP}$, $V_{F,P}$ and cov_F been the same as some DGP from $\mathcal{F}$ and had the asymptotic bias d been infinite. Note that if $u_{F,d} \in \mathcal{U}_\infty$ (i.e., $\|d\| = \infty$), the corresponding δ approaches to infinity at faster than $n^{1/2}$ rate, corresponding to *severely* misspecification of the parametric restrictions. Together, $\mathcal{U}$ and $\mathcal{U}_\infty$ are a device that allows me to compare the asymptotic risk of $\hat{\beta}_{n,\hat{w}_n}$ to that of $\hat{\beta}_{n,SP}$ uniformly over varied degrees of misspecification of the parametric restrictions.

To show the main uniform dominance result, I will first approximate the bounds of asymptotic risk difference using transformation of $r(u)$ for the mildly misspecified case (including the correct specification case, encompassed in $\mathcal{U}$) and the severely misspecified case separately (encompassed in $\mathcal{U}_\infty$), and then combine the two cases together.

Lemma 2. *Suppose: (i) Conditions 1 - 3 hold; (ii) $tr(A_F) > 0$ and $tr(B_F) > 0$. Then*

$$Asy\overline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}), 0\right\} \tag{4.15}$$

$$Asy\underline{RD}(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d}), 0\right\}. \tag{4.16}$$

Proof. See Appendix B. □

If the parametric restrictions are severely misspecified, then I have $u_{F,d} \in \mathcal{U}_\infty$ (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 $\inf_{u_{F,d}\in\mathcal{U}_\infty} r(u_{F,d}) = 0$). As mentioned before, the former corresponds to the DGPs under which the parametric restrictions are correctly specified or mildly misspecified, and the later corresponds to the severely misspecified case. $\max\left\{\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}), 0\right\}$ characterizes the least favorable DGP for the averaging estimator, and $\min\left\{\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d}), 0\right\}$ characterizes the most favorable.

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 uniform dominance result.

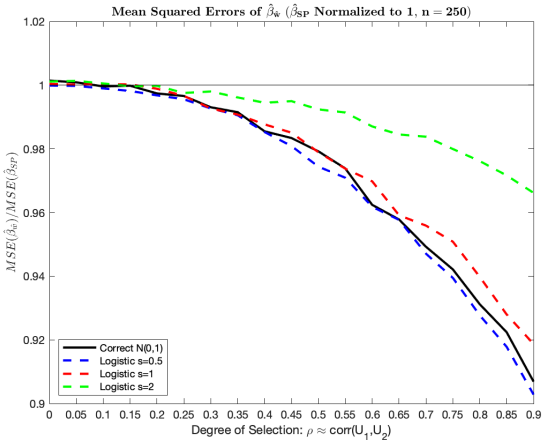
Theorem 1. *Suppose Conditions 1 - 3 hold. If* $tr(A_F) > 0$, $tr(B_F) > 0$ *and* $tr(A_F) \geq 4\rho_{\max}(A_F)$ *for any* $F \in \mathcal{F}$ *with* $\delta_F = 0$, *then equation* (4.5) *holds. If, in addition,* $tr(A_F) > 4\rho_{\max}(A_F)$, *then equation* (4.4) *holds. Thus, the averaging estimator* $\hat{\beta}_{n,\hat{w}_n}$ *uniformly dominates the semiparametric estimator* $\hat{\beta}_{n,SP}$.

Proof. See Appendix B. □

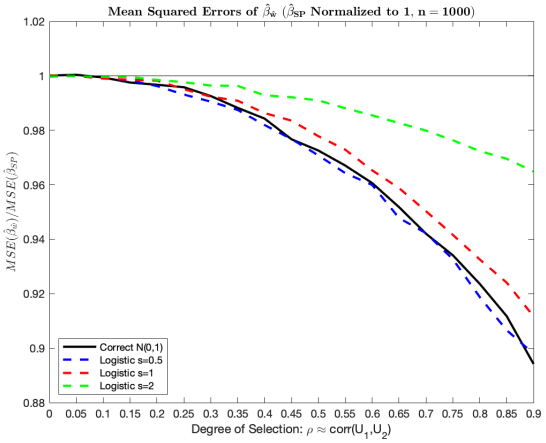
To give some intuition for the conditions in Theorem 1, let us consider the case where the researcher chooses $\Upsilon = (V_{F,SP} - cov_F)^{-1}$. In this case, the condition $\text{tr}(A_F) > 0$ becomes $V_{F,SP} > cov_F$, which is a necessary condition for $V_{F,SP} > V_{F,P}$. The latter indicates that the parametric estimator should achieve strict efficiency gain over the semiparametric estimator. And 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 dimension condition is common for shrinkage estimators. For example, my 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 averaging estimator in Cheng, Liao, and Shi (2019), and is weaker than $k \geq 5$ for the the estimators in Judge and Mittelhammer (2004) and Mittelhammer and Judge (2005).

5 Monte Carlo Experiments

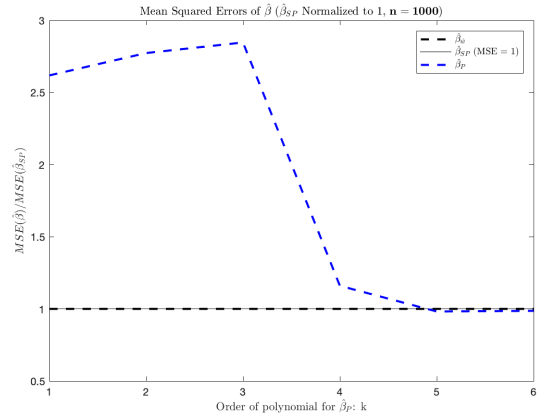
Example 1 (cont'd) - Sample Selection Model. *In my Monte Carlo experiments, I let* $\beta_1 = (1, 2, 3, 4)'$, $\beta_2 = (1, 1, 1, 1)'$, *and* $(X_{1i}', X_{2i}')' \sim \mathcal{N}(0_{8\times 1}, I_8)$. *The first step parametric restriction is that the two error terms* (U_{1i}, U_{2i}) *are jointly normally distributed. For the correct specification, I let the true marginal distributions of* U_{1i} *and* U_{2i} *be both standard normal. For the misspecification, I let them be logistic distributions with mean* $\mu = 0$ *and variance* $s^2\pi^2/3$, *and vary the values* $s \in \{0.5, 1, 2\}$ *(hence also vary the degree of misspecification). For both the correct specification and the misspecification, I vary the degree of selection by varying* $\rho = \text{Corr}(U_{1i}, U_{2i})$ *from* 0 *to* 0.9 *with* 0.05 *steps. Note that when* $\rho = 0$, *there is no misspecification in all of these cases since there is no sample selection*



(a) Example 1: Sample Selection Model ($n = 250$)



(b) Example 1: Sample Selection Model ($n = 1000$)



(c) Example 2: Quantile Treatment Effects Model ($n = 1000$)

Figure 1: Simulation Results for Examples 1 and 2

(so no correction is needed). Different sample sizes $n \in \{250, 1000\}$ *are considered, and I repeat* $R = 10000$ *times. The semiparametric estimator* $\hat{\beta}_{n,SP}$ *uses the kernel approximation for the h function, and the parametric estimator* $\hat{\beta}_{n,P}$ *first estimates* β_2 *using probit, then plugs in the inverse Mill's ratio to estimate* β_1*. I choose* $\Upsilon = I_4$*, and normalize the MSE of the semiparametric estimator to unity.*

The normalized MSEs of the averaging estimator are reported in Figures 1(a) and 1(b), separately for different sample sizes. The thin black line at the level of one is the semiparametric benchmark, being below it means that the averaging estimator has smaller MSE than the semiparametric estimator. The thick black line represents the normalized MSE of the averaging estimator under the correct specification. The blue, red and green dashed lines represent normalized MSEs of the averaging estimator when the degree of misspecification s is $0.5, 1, 2$*, respectively.*

A few observations can be made here. First, when the normality restriction is correctly specified, the averaging estimator performs very well. This is natural, since in this case the parametric estimator is consistent and more efficient. Second, when the normality restriction is misspecified, the averaging estimator still has smaller MSEs than the robust semiparametric estimator, regardless of the degree of misspecification and the degree of selection. Overall, the averaging estimator achieves sizable improvement compared to the semiparametric benchmark, and the averaging weight is strictly between zero and one, because the efficiency gain of $\hat{\beta}_{n,P}$ *outweighs its bias in this particular Monte Carlo setting.*

Example 2 (cont'd) - Quantile Treatment Effects. *I modify the Monte Carlo model used by Firpo (2007) and focus on the quantiles at* $\tau = (0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95)$*. The parameter values are chosen such that the true QTEs are* $\Delta = (7.59, 6.76, 5.64, 5.00, 4.36, 3.24, 2.41)$.[19] *I consider sample size* $n = 1000$ *and repeat* $R = 1000$ *times. The semiparametric estimator* $\hat{\beta}_{n,SP}$ *uses an 8th order polynomial logit in* (X_{1i}, X_{2i}) *proposed by Hirano, Imbens, and Ridder (2003) (and also employed by Firpo, 2007) to estimate the propensity score; the parametric estimator* $\hat{\beta}_{n,P}$ *uses polynomial probit in* (X_{1i}, X_{2i}) *to estimate the propensity score, and I vary the order of the polynomial* $k \in \{1, \ldots, 6\}$*. I choose* $\Upsilon = I_7$ *and normalize the MSE of the semiparametric estimator to unity.*

Figure 1(c) plots the MSEs of the averaging estimator and of the parametric estimator for different polynomial probit order k. The thin black line at the level of one is again the semiparametric benchmark. The blue dashed line represents the MSEs of the parametric estimator. In this particular setting, all of the polynomial probit are misspecified,[20] *and the misspecification bias outweighs the efficiency gain by imposing a parametric restriction. As*

[19] See Section 2 of the supplement to Firpo (2007) for details of the model. The parameter values I choose are $\mu_1 = 1$, $\mu_2 = 5$, $\delta_0 = -1$, $\delta_1 = 5$, $\delta_2 = -1$, $\delta_3 = -0.05$, $\gamma_1 = -5$, $\gamma_2 = 1$, $\beta = 5$, $\sigma_{\epsilon_0} = 5$ and $\sigma_{\epsilon_1} = 2.5$.

[20] In fact, the propensity score is a logit function that depends on a quadratic form of (X_{1i}, X_{2i}). See Section 2 of the supplement to Firpo (2007) for details.

a result, none of the MSEs of the parametric estimator $\hat{\beta}_{n,P}$ is below the benchmark. For smaller k, the MSEs of $\hat{\beta}_{n,P}$ can be as large as nearly three times of those of $\hat{\beta}_{n,SP}$. On the contrary, the MSEs of the averaging estimator $\hat{\beta}_{\hat{w}}$ (black dashed line) never go higher than those of $\hat{\beta}_{n,SP}$, even for smaller k. This experiment highlights the robustness of the averaging estimator.

Unlike Example 1, the averaging estimator performs the same as $\hat{\beta}_{n,SP}$ (dashed and solid black lines on top of each other), and the averaging weight is almost always zero, because the performance of $\hat{\beta}_{n,P}$ (dashed blue line) is much worse than $\hat{\beta}_{n,SP}$ in this particular Monte Carlo setting.

6 Conclusion

This paper studies the two step M estimation of a finite dimensional parameter in a semiparametric model which contains a potentially infinite dimensional first step nuisance parameter. I present an averaging estimator that combines a semiparametric estimator based on nonparametric first step and a parametric estimator which imposes parametric restrictions on the first step, where the averaging weight is the sample analog of an infeasible optimal weight that minimizes quadratic risk functions. Using a uniform asymptotic framework, I show that under mild sufficient conditions, the asymptotic lower bound of the truncated quadratic risk differences between the averaging estimator and the semiparametric estimator is strictly less than zero under a class of DGPs that includes both correct specification and misspecification of the parametric restrictions, and the asymptotic upper bound is weakly less than zero. This uniform dominance of the averaging estimator is illustrated by two specific widely used semiparametric models.

Appendix A Details on the Examples

Example 1 (cont'd) - Sample Selection Model. *To focus on the sample selection issue, I assume that* $\mathbb{E}_F(U_{1i}|X_{1i}, X_{2i}) = \mathbb{E}_F(U_{2i}|X_{1i}, X_{2i}) = 0$, $\mathbb{E}_F(U_{1i}^2|X_{1i}, X_{2i}) = \mathbb{E}_F(U_{2i}^2|X_{1i}, X_{2i}) = 1$, *and let* $\rho_F \equiv \mathbb{E}_F(U_{1i}U_{2i}|X_{1i}, X_{2i})$ *for all DGPs in* $\mathcal{F}$. *Let* $\lambda_i \equiv \lambda(-X_{2i}'\gamma_F)$ *denote the inverse Mill's ratio. Under the joint normality restriction,* V_{1i} *is conditionally heteroskedastic, i.e.,* $\mathbb{E}_F(V_{1i}^2|X_{1i}, \lambda_i, Y_{2i} = 1) = (1-\rho_F^2)+\rho_F^2(1-\lambda_i X_{2i}'\gamma_F-\lambda_i^2)$, *so it is tempting to use feasible generalized least square (FGLS) in the second step of Heckman (1979) two step estimator. Heckman (1979), however, points out that the FGLS estimator is not asymptotically efficient due to non-diagonality of the information matrix (see Heckman, 1977, for details), and the MLE possesses undesirable numerical properties (see Wales and Woodland, 1980; Nelson, 1984, for more discussion). Since the proposed averaging estimator does not require* $\hat{\beta}_{n,P}$ *to be asymptotically efficiency under the parametric restriction in order to achieve dominance, I employ OLS in the second step of* $\hat{\beta}_{n,P}$ *in this example for exposition purpose. For* $\hat{\beta}_{n,SP}$, *I follow Robinson (1988) and use the Nadaraya-Watson estimators for the nuisance functions* $h_{1F}(x_2) \equiv \mathbb{E}_F(Y_1|X_2 = x_2)$ *and* $h_{2F}(x_2) \equiv \mathbb{E}_F(X_1|X_2 = x_2)$.

In this example, let $Q(z, \beta, h) \equiv \frac{1}{2}[y_1 - h_1(x_2) - (x_1 - h_2(x_2))'\beta]^2$ *where* z *represents the vector of all observed variables, so that* $Q_F(\beta, h)$ *in equation* (3.1) *equals to* $\mathbb{E}_F[Q(Z, \beta, h)]$, *where the expectation is taken with regard to the data* Z. *Borrowing some notation from Ichimura and Lee (2010) and noting that* h *does not depend on* β, *one gets*

$$
\begin{aligned}
\Delta_1(z) \equiv D_\beta Q(z, \beta, h) &= [y_1 - h_1(x_2) - (x_1 - h_2(x_2))'\beta](x_1 - h_2(x_2)), \\
D_{\beta\beta'}Q(z, \beta, h) &= (x_1 - h_2(x_2))(x_1 - h_2(x_2))', \\
V_0 = \frac{d^2Q(\beta, h)}{d\beta d\beta'} &= D_{\beta\beta'}Q(\beta, h) = \mathbb{E}[(X_1 - h_2(X_2))(X_1 - h_2(X_2))'], \\
D_hQ(z, \beta, h_F)[h] &= -[y_1 - h_{1F}(x_2) - (x_1 - h_{2F}(x_2))'\beta](h_1(x_2) - h_2(x_2)'\beta), \\
\frac{d}{d\beta'}D_hQ(\beta, h_F)[h] &= D_{\beta h}Q(\beta, h_F)[h] \\
&= \mathbb{E}_F[(X_1 - h_{2F}(X_2))'\beta(h_1(X_2) - h_2(X_2)'\beta] \\
&\quad + \mathbb{E}_F[(Y_1 - h_{1F}(X_2) - (X_1 - h_{2F}(X_2))'\beta)h_2(X_2)] \\
&= 0,
\end{aligned}
$$

where the last equality holds by the law of iterated expectations (i.e., first conditional on X_2*). This implies that* $\Gamma_1(z)$, *the term in Ichimura and Lee (2010) that captures the impact of first step estimation errors of* h *on the asymptotic distribution of* $\hat{\beta}_n$, *is zero. By Theorem 3.3 of Ichimura and Lee (2010), the non-centered influence function of an estimator* $\hat{\beta}_n$ *is* $\psi(z) = V_0^{-1}\Delta_1(z)$. *The sample analogs of the non-centered influence functions of* $\hat{\beta}_{n,SP}$ *and*

$\hat{\beta}_{n,P}$ *in equations* (3.13) *and* (3.14)*, therefore, follow this general formula.*

Example 2 (cont'd) - Quantile Treatment Effects. *First I describe how to obtain the estimator* $\hat{\beta}_n$ *once the estimated propensity score function* $\hat{p}_n(x)$ *is available. Let* $U_i \equiv \max\{Y_i - q, 0\}$*,* $V_i \equiv \max\{q - Y_i, 0\}$*; and define the* $n \times 1$ *vectors* $Y \equiv (Y_1, \ldots, Y_n)'$*,* $U \equiv (U_1, \ldots, U_n)'$*,* $V \equiv (V_1, \ldots, V_n)'$*,* $\hat{w}_{n,1} \equiv (\hat{w}_{n,1,1}, \ldots, \hat{w}_{n,1,n})'$ *and* $\hat{w}_{n,0} \equiv (\hat{w}_{n,0,1}, \ldots, \hat{w}_{n,0,n})'$*, where* $\hat{w}_{n,j,i}$ *(*$j = 0, 1$ *and* $i = 1, \ldots, n$*) are defined in equation (2.5). For fixed* $\tau \in (0, 1)$*, the following linear programming problem gives rise to* $\hat{q}_{1,\tau}$*:*

$$\begin{aligned} \min_{(q,U',V')'} & \tau \hat{w}'_{n,1} U + (1-\tau) \hat{w}'_{n,1} V \\ s.t. & q\mathbb{I} + U - V - Y = 0, \\ & (q, U', V')' \in \mathbb{R} \times \mathbb{R}^{2n}_+; \end{aligned}$$

and the following linear programming problem gives rise to $\hat{q}_{0,\tau}$*:*

$$\begin{aligned} \min_{(q,U',V')'} & \tau \hat{w}'_{n,0} U + (1-\tau) \hat{w}'_{n,0} V \\ s.t. & q\mathbb{I} + U - V - Y = 0, \\ & (q, U', V')' \in \mathbb{R} \times \mathbb{R}^{2n}_+. \end{aligned}$$

Next, one may follow Firpo (2007) to use the series logit propensity score estimator as $\hat{p}_n(x)$ *and get* $\hat{\beta}_{n,SP}$*. Alternatively, one may use the linear probit propensity score estimator* $\Phi(x'\hat{\gamma}_n)$ *as* $\hat{p}(x)$ *to get* $\hat{\beta}_{n,P}$*.*

Theorem 1 in Firpo (2007) shows that the influence function of $\hat{\beta}_{n,SP}$ *is given by* $\psi_\tau(y, t, x)$ *in the following equation (A.1).*

$$\begin{aligned} \pi_{j,\tau}(y) &\equiv -\frac{\mathbb{I}\{y \leq q_{j,\tau}\} - \tau}{f_j(q_{j,\tau})},\ j = 0, 1, \\ \psi_{1,\tau}(y, t, x) &\equiv \frac{t}{p(x)} \cdot \pi_{1,\tau}(y) - \frac{t - p(x)}{p(x)} \cdot \mathbb{E}[\pi_{1,\tau}(Y)|x, T = 1], \\ \psi_{0,\tau}(y, t, x) &\equiv \frac{1-t}{1-p(x)} \cdot \pi_{0,\tau}(y) + \frac{t - p(x)}{1-p(x)} \cdot \mathbb{E}[\pi_{0,\tau}(Y)|x, T = 0], \\ \psi_\tau(y, t, x) &\equiv \psi_{1,\tau}(y, t, x) - \psi_{0,\tau}(y, t, x), \end{aligned} \tag{A.1}$$

where $f_j(q)$ *denotes the probability density (pdf) function of the potential outcome* $Y_{j,i}$ *evaluated at* $q \in \mathbb{R}$ *for* $j = 0, 1$*. By slightly modifying the argument in the proof of Theorem 1 in Firpo (2007) (in Appendix 3 of that paper), one can show that the influence function of*

$\hat{\beta}_{n,P}$ is given by $\psi_{\tau,P}(y,t,x)$ in equation (A.2) as follows.[21]

$$\begin{aligned}
\pi_{j,\tau,P}(y) &\equiv -\frac{\mathbb{I}\{y \le q_{j,\tau,P}\} - \tau}{f_j(q_{j,\tau,P})},\ j = 0,1, \\
\psi_{1,\tau,P}(y,t,x) &\equiv \frac{t}{\Phi(x'\gamma)} \cdot \pi_{1,\tau,P}(y) - \frac{t - \Phi(x'\gamma)}{\Phi(x'\gamma)} \cdot \mathbb{E}[\pi_{1,\tau,P}(Y)|x, T=1], \\
\psi_{0,\tau,P}(y,t,x) &\equiv \frac{1-t}{1-\Phi(x'\gamma)} \cdot \pi_{0,\tau,P}(y) + \frac{t - \Phi(x'\gamma)}{1-\Phi(x'\gamma)} \cdot \mathbb{E}[\pi_{0,\tau,P}(Y)|x, T=0], \\
\psi_{\tau,P}(y,t,x) &\equiv \psi_{1,\tau,P}(y,t,x) - \psi_{0,\tau,P}(y,t,x),
\end{aligned} \tag{A.2}$$

where $q_{j,\tau,P}$ denotes the pseudo-true quantile under the probit restriction. Note that in the above equation, f_j and $\mathbb{E}$ denote the true pdf function and expectation operator, regardless of the probit restriction. The sample analogs of these influence functions, therefore, are given in equations (3.15) *and* (3.16)*.*

Appendix B Proofs

Justification for Condition 2(ii).

Note that the asymptotic properties of $\hat{\beta}_{n,SP}$ do not depend on whether $\|d\| < \infty$ or $\|d\| = \infty$, so we still have $n^{1/2}(\hat{\beta}_{n,SP} - \beta_{F_n}) \xrightarrow{d.} \xi_{F,SP}$ under the same low-level conditions like in part (i).

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

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

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

For case (ii), note that $\beta_{F,P}$ is defined in equation (4.7), then under the same conditions for $\hat{\beta}_{n,SP} = \beta_{F_n} + o_p(1)$, one gets $\hat{\beta}_{n,P} = \beta_{F_n,P} + o_p(1)$.[22] This, combined with the presumption

[21] Firpo (2007) approach is in turn based on Hirano, Imbens, and Ridder (2003) and Newey (1995). Details are omitted here.

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

that $\|\delta_{F_n}\| = \|\beta_{F_n,P} - \beta_{F_n}\| > c$, implies that

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

Proof of Lemma 1

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

Part (ii). Because the probability limits of $\hat{V}_{n,SP}$, $\hat{V}_{n,P}$ and $\widehat{cov}_n$ 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 the Slutsky's theorem, implies that $n^{1/2}(\hat{\beta}_{n,\hat{w}_n} - \beta_F) \xrightarrow{d.} \xi_{F,SP}$.

□

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

$$R_\zeta(u_{F,d}) \equiv \mathbb{E}_F\left(\min\left\{\xi'_{F,SP}\Upsilon\xi_{F,SP}, \zeta\right\}\right), \tag{B.2}$$

$$\bar{R}_\zeta(u_{F,d}) \equiv \begin{cases} \mathbb{E}_F\left(\min\left\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}, \zeta\right\}\right), & \text{if } \|d\| < \infty, \\ \mathbb{E}_F\left(\min\left\{\xi'_{F,SP}\Upsilon\xi_{F,SP}, \zeta\right\}\right), & \text{if } \|d\| = \infty, \end{cases} \tag{B.3}$$

$$r_\zeta(u_{F,d}) \equiv \bar{R}_\zeta(u_{F,d}) - R_\zeta(u_{F,d}) \tag{B.4}$$

$$= \begin{cases} \mathbb{E}_F\left(\min\{\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}, \zeta\} - \min\{\xi'_{F,SP}\Upsilon\xi_{F,SP}, \zeta\}\right), & \text{if } \|d\| < \infty, \\ 0, & \text{if } \|d\| = \infty, \end{cases} \tag{B.5}$$

$$r(u_{F,d}) \equiv \begin{cases} \mathbb{E}_F\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} - \xi'_{F,SP}\Upsilon\xi_{F,SP}\right), & \text{if } \|d\| < \infty, \\ 0, & \text{if } \|d\| = \infty. \end{cases} \tag{B.6}$$

For any positive finite ζ, define

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \equiv \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{E}_F[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_F) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_F)], \tag{B.7}$$

$$Asy\underline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \equiv \liminf_{n\to\infty} \inf_{F\in\mathcal{F}} \mathbb{E}_F[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_F) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_F)]. \tag{B.8}$$

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

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \leq \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}, \tag{B.9}$$

$$Asy\underline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \geq \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}. \tag{B.10}$$

Proof. First I prove inequality (B.9). By the definition of supremum and the definition of $Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP})$ in equation (B.7), there exists a sequence of DGPs, denoted by $\{F_n\}_{n\in\mathbb{N}}$, such that

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

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

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \lim_{n\to\infty} \mathbb{E}_{F_{p_n}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})]. \tag{B.11}$$

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

$$p_n^{*1/2}\delta_{F_{p_n^*}} \to d \text{ for some } d \in \mathbb{R}_\infty \text{ and } S(F_{p_n^*}) \longrightarrow S(F^*) \text{ for some } F^* \in \mathcal{F}. \tag{B.12}$$

Note that for any subsequence of $\{p_n\}_{n\in\mathbb{N}}$, the limit of the right hand side in equation (B.11) remains the same, which implies

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \lim_{n\to\infty} \mathbb{E}_{F_{p_n^*}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n^*}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n^*}})]. \tag{B.13}$$

Equation (B.4) suggests that in order to prove equation (B.7), one needs to link the right hand side of equation (B.13) with $R_\zeta(u_{F,d})$ and $\bar{R}_\zeta(u_{F,d})$ defined in equations (B.2) and (B.3). First consider the case where $\|d\| < \infty$ in equation (B.12). By Condition 2(i) and

Lemma 1(i),

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

which combined with the continuous mapping theorem implies that

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

Since Υ is positive semi-definite, $\xi'_{F,SP}\Upsilon\xi_{F,SP}$ and $\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}$ are both nonnegative. Note that the function $f(x) \equiv \min\{x, \zeta\}$ is a bounded continuous function of $x \geq 0$ for fixed positive ζ. Applying the Portmanteau lemma (e.g., Lemma 2.2 in Van der Vaart, 2000) and invoking equations (B.2) and (B.3), one gets

$$\mathbb{E}_{F_{p_n^*}}\left[\ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n^*}})\right] \to R_\zeta(u_{F^*,d}) \text{ and } \mathbb{E}_{F_{p_n^*}}\left[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n^*}})\right] \to \bar{R}_\zeta(u_{F^*,d}). \tag{B.14}$$

Then consider the case where $\|d\| = \infty$ in equation (B.12). By Condition 2(ii) and Lemma 1(ii),

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

Using the same argument, one also gets equation (B.14) in this case. Combine equations (B.4), (B.13) and (B.14), one can unify the two cases and write

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

This proves equation (B.9).

The proof of equation (B.10) follows the same argument and hence is omitted here. □

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

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

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

Proof. First I prove inequality (B.15). By the definition of $\mathcal{U}$ in equation (4.13), $\delta_F = 0$ for any $F \in \mathcal{F}$ such that $u_{F,d} \in \mathcal{U}$. For any $u_{F,d} \in \mathcal{U}$, let N_{ϵ_F} denote the smallest n such

that $n^{-1/2}\|d\| < \epsilon_F$, where ϵ_F satisfies Condition 3(ii). Then by Condition 3(ii), for each $n \geq N_{\epsilon_F}$, there is an $F_n \in \mathcal{F}$ with $\delta_{F_n} = n^{-1/2}\|d\|$ and $\|\bar{S}(F_n) - \bar{S}(F)\| \leq n^{-1/2}C\|d\|^\kappa$ for some $C, \kappa > 0$. For each $n \leq N_{\epsilon_F}$, let $F_n = F$. Therefore, a sequence of DGPs $\{F_n\}_{n\in\mathbb{N}}$ in $\mathcal{F}$ satisfying $n^{1/2}\delta_{F_n} \to d$ and $\bar{S}(F_n) \to \bar{S}(F)$ is constructed for any $u_{F,d} \in \mathcal{U}$. Recalling the definition of $\bar{S}(F)$ after equation (4.8), this immediately implies that for such $\{F_n\}_{n\in\mathbb{N}}$,

$$n^{1/2}\delta_{F_n} \to d \in \mathbb{R}^k,\ V_{F_n,SP} \to V_{F,SP},\ cov_{F_n} \to cov_F, \text{ and } V_{F_n,P} \to V_{F,P}. \tag{B.17}$$

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

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})] = \limsup_{n\to\infty} \mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})]. \tag{B.18}$$

Since $\{p_n\}_{n\in\mathbb{N}}$ is a subsequence of $\{n\}_{n\in\mathbb{N}}$, equation (B.17) implies that

$$n^{1/2}\delta_{F_{p_n}} \to d \in \mathbb{R}^k,\ V_{F_{p_n},SP} \to V_{F,SP},\ cov_{F_{p_n}} \to cov_F, \text{ and } V_{F_{p_n},P} \to V_{F,P}. \tag{B.19}$$

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

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

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

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\left[\ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_{p_n}})\right] = R(u_{F,d}), \text{ and } \lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\left[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_{p_n}})\right] = \bar{R}(u_{F,d}). \tag{B.20}$$

This, combined with equation (B.18), the definition of $Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP})$ in equation (B.7) and the definition of $r(u_{F,d})$ in equation (B.4), implies that for any $u_{F,d} \in \mathcal{U}$,

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

which further implies that

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \geq \sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}). \tag{B.21}$$

On the other hand, by the definition of $\mathcal{U}_\infty$ in equation (4.14), for any $u_{F,d} \in \mathcal{U}_\infty$, either (i) $\delta_F = 0$ or (ii) $\|\delta_F\| > 0$. For case (i), let $\mathbb{I}_k$ be a $k \times 1$ vector of ones and let N_{ϵ_F} denote the smallest n such that $n^{-1/4}\|\mathbb{I}_k\|^{1/2} = n^{-1/4}k^{1/2} < \epsilon_F$, where ϵ_F satisfies Condition 3(ii).

Then by Condition 3(ii), for each $n \geq N_{\epsilon_F}$, there is an $F_n \in \mathcal{F}$ with $\delta_{F_n} = n^{-1/4}k^{1/2}$ and $\|\bar{S}(F_n) - \bar{S}(F)\| \leq Cn^{-\kappa/4}k^{\kappa/2}$ for some $C, \kappa > 0$. For each $n \leq N_{\epsilon_F}$, let $F_n = F$. For case (ii), let $F_n = F$ for $n = 1, 2, \ldots$. Therefore, a sequence of DGPs $\{F_n\}_{n\in\mathbb{N}}$ in $\mathcal{F}$ satisfying $n^{1/2}\delta_{F_n} \to \infty$, $\delta_{F_n} \to F$ and $\bar{S}(F_n) \to \bar{S}(F)$ is constructed for any $u_{F,d} \in \mathcal{U}_\infty$, regardless of whether $\delta_F = 0$ or $\|\delta_F\| > 0$. Recalling the definition of $\bar{S}(F)$ after equation (4.8), this immediately implies that for such $\{F_n\}_{n\in\mathbb{N}}$,

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

Again, recalling the definition of $\bar{S}(F)$ after equation (4.8), this immediately implies that for the $\{F_n\}_{n\in\mathbb{N}}$ in $\mathcal{F}$,

$$n^{1/2}\delta_{F_n} \to d \in \mathbb{R}^k,\ V_{F_n,SP} \to V_{F,SP},\ cov_{F_n} \to cov_F, \text{ and } V_{F_n,P} \to V_{F,P}.$$

Then similar argument used to show equations (B.18) - (B.20) can be applied to show that there exists a subsequence $\{p_n\}_{n\in\mathbb{N}}$ of $\{n\}_{n\in\mathbb{N}}$ such that equations (B.18) and (B.20) are satisfied, with the help of Condition 2(ii) and Lemma 1(ii). Combining this with the definition of $Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP})$ in equation (B.7) and the definition of $r(u_{F,d})$ in equation (B.4), implies that for any $u_{F,d} \in \mathcal{U}_\infty$,

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) \geq \limsup_{n\to\infty} \mathbb{E}_{F_n}[\ell_\zeta(\hat{\beta}_{n,\hat{w}_n}, \beta_{F_n}) - \ell_\zeta(\hat{\beta}_{n,SP}, \beta_{F_n})] = 0. \tag{B.22}$$

Equation (B.15) immediately follows inequalities (B.21) and (B.22).

The proof of equation (B.16) follows the same argument and hence is omitted here. □

Lemma B.3. *Suppose: (i) Conditions 1 - 3 hold; (ii) $tr(A_F) > 0$ and $tr(B_F) > 0$. Then*

$$\sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left(\xi'_{F,SP}\Upsilon\xi_{F,SP}\right)^2\right] \leq C, \tag{B.23}$$

$$\sup_{u_{F,d}\in\mathcal{U}} \mathbb{E}\left[\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}\right)^2\right] \leq C. \tag{B.24}$$

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

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

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

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

This implies that

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

where the second inequality holds because $\mathcal{Z}\sim\mathcal{N}\left(0_{k\times 1}, I_{k\times k}\right)$. Equation (B.23) follows since the upper bound does not depend on F.

By the definition of $\bar{\xi}_{F,d}$, the Cauchy-Schwarz inequality and the simple inequality $2|ab| \leq a^2+b^2$ for any real numbers a and b, one

$$\begin{aligned}\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d} &\leq 2\xi'_{F,SP}\Upsilon\xi_{F,SP} + 2w^2_F\left(\xi_{F,P}+d-\xi_{F,SP}\right)'\Upsilon\left(\xi_{F,P}+d-\xi_{F,SP}\right)\\ &= 2\xi'_{F,SP}\Upsilon\xi_{F,SP} + 2w^2_F\left(\tilde{\xi}_F+\tilde{d}_F\right)'D\left(\tilde{\xi}_F+\tilde{d}_F\right),\end{aligned} \tag{B.25}$$

where

$$D \equiv [\ -I_k \quad I_k\ ]'\Upsilon[\ -I_k \quad I_k\ ] \tag{B.26}$$

to facilitate the analysis. Combine equation (B.25) and the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$ for any real numbers a and b, one gets

$$\begin{aligned}\left(\bar{\xi}'_{F,d}\Upsilon\bar{\xi}_{F,d}\right)^2 \leq& 8\left(\xi'_{F,SP}\Upsilon\xi_{F,SP}\right)^2 + 8\left[w^2_F\left(\tilde{\xi}_F+\tilde{d}_F\right)'D\left(\tilde{\xi}_F+\tilde{d}_F\right)\right]^2\\ \leq& C + 8\left[w^2_F\left(\tilde{\xi}_F+\tilde{d}_F\right)'D\left(\tilde{\xi}_F+\tilde{d}_F\right)\right]^2,\end{aligned} \tag{B.27}$$

where the second inequality is by equation (B.23). By the definitions of w_F in Lemma 1 and of A_F and B_F in Theorem 1, one has

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

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

$$\mathbb{E}\left[w^2_F\left(\tilde{\xi}_F+\tilde{d}_F\right)'D\left(\tilde{\xi}_F+\tilde{d}_F\right)\right]^2 \leq 2C[\mathrm{tr}(\Upsilon V_{F,SP})]^2 + 2C[\mathrm{tr}(\Upsilon cov_F)]^2$$

$$\leq 2C[\text{tr}(\Upsilon V_{F,SP})]^2 + 2C[\text{tr}(\Upsilon V_{F,SP})]^2 \leq C, \qquad \text{(B.28)}$$

where the second inequality holds by Condition 2(i), Condition 3(i) and that the Cauchy-Schwarz inequality implies $cov_F \leq \max\{V_{F,SP}, V_{F,P}\}$ for any $F \in \mathcal{F}$. Together, equations (B.27) and (B.28) imply equation (B.24), since the upper bound does not depend on F. □

Lemma B.4. *Suppose: (i) Conditions 1 - 3 hold; (ii) $tr(A_F) > 0$ and $tr(B_F) > 0$. Then*

$$\lim_{\zeta \to \infty} \sup_{u_{F,d} \in \mathcal{U}} |r_\zeta(u_{F,d}) - r(u_{F,d})| = 0. \qquad \text{(B.29)}$$

Proof. First note that

$$\begin{aligned}
&\sup_{u_{F,d} \in \mathcal{U}} \left| \mathbb{E}\left[\min\{\bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d}, \zeta\} - \bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} \right] \right| \\
=& \sup_{u_{F,d} \in \mathcal{U}} \left| \mathbb{E}\left[\left(\zeta - \bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d}\right) \mathbb{I}\left\{ \bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} > \zeta \right\} \right] \right| \\
\leq& \sup_{u_{F,d} \in \mathcal{U}} \mathbb{E}\left[\left| \zeta - \bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} \right| \cdot \mathbb{I}\left\{ \bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} > \zeta \right\} \right] \\
\leq& \zeta \sup_{u_{F,d} \in \mathcal{U}} \mathbb{E}\left[\mathbb{I}\left\{ \bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} > \zeta \right\} \right] + \sup_{u_{F,d} \in \mathcal{U}} \mathbb{E}\left[\left(\bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d}\right) \cdot \mathbb{I}\left\{ \bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} > \zeta \right\} \right] \\
\leq& 2\zeta^{-1} \sup_{u_{F,d} \in \mathcal{U}} \mathbb{E}\left[\left(\bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d}\right)^2 \right] \leq 2C\zeta^{-1}, \qquad \text{(B.30)}
\end{aligned}$$

where the first equality is by the fact that $\min\{x, \zeta\} - x = (\zeta - x) \cdot \mathbb{I}\{x > \zeta\}$; the first inequality is by the Jensen's inequality and the fact that an indicator function is always non-negative; the second inequality holds because $\zeta > 0$, $\bar{\xi}'_{F,d} \Upsilon \bar{\xi}_{F,d} \geq 0$, and the simple inequality $|a - b| \leq a + b$ for any non-negative real numbers a and b; the third inequality holds by the Markov's inequality;[23] the fourth inequality is by (B.24).

By (B.23) and the same argument, I can show that

$$\sup_{u_{F,d} \in \mathcal{U}} |\mathbb{E}\left[\min\{\xi_{F,SP} \Upsilon \xi_{F,SP}, \zeta\} - \xi_{F,SP} \Upsilon \xi_{F,SP} \right]| \leq 2C\zeta^{-1}. \qquad \text{(B.31)}$$

Combining inequalities (B.30) and (B.31), the definitions of $r_\zeta(u_{F,d})$ and $r(u_{F,d})$ in equations (B.5) and (B.6), and the triangular inequality, one gets $\sup_{u_{F,d} \in \mathcal{U}} |r_\zeta(u_{F,d}) - r(u_{F,d})| \leq 4C\zeta^{-1}$, which immediately implies equation (B.29). □

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

Proof of Lemma 2

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

$$Asy\overline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}, \tag{B.32}$$

$$Asy\underline{RD}_\zeta(\hat{\beta}_{n,\hat{w}_n}, \hat{\beta}_{n,SP}) = \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\}, \tag{B.33}$$

for any finite $\zeta > 0$. Then note that Lemma B.4 implies

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

Moreover, note that for $u_{F,d} \in \mathcal{U}_\infty$, equations (B.5) and (B.2) imply that both $r_\zeta(u_{F,d}) = r(u_{F,d}) = 0$. Furthermore, since $\max\{x, 0\}$ and $\min\{x, 0\}$ are both continuous functions of x, the equalities above remain valid after applying these continuous functions; that is,

$$\lim_{\zeta\to\infty} \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\} = \max\left\{\sup_{u_{F,d}\in\mathcal{U}} r(u_{F,d}), 0\right\}, \tag{B.34}$$

$$\lim_{\zeta\to\infty} \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r_\zeta(u_{F,d}), 0\right\} = \min\left\{\inf_{u_{F,d}\in\mathcal{U}} r(u_{F,d}), 0\right\}. \tag{B.35}$$

Combining equations (B.32), (B.34) and the definitions in equations (4.3) and (B.7) gives the result in equation (4.15). Combining equations (B.33), (B.35) and the definitions in equations (4.2) and (B.8) gives the result in equation (4.16). □

Proof of Theorem 1

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

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

By the definitions of w_F in equation (4.10) and of A_F and B_F in Lemma 1, this implies that for any $u_{F,d} \in \mathcal{U}$,

$$r(u_{F,d}) = 2\text{tr}(A_F)J_{1,F} + [\text{tr}(A_F)]^2 J_{2,F}, \tag{B.36}$$

where

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

Define

$$E \equiv [\ -I_k \quad I_k\ ]'\Upsilon[\ I_k \quad 0_{k\times k}\ ],$$

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

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

where $\tilde{\xi}_F$ and $\tilde{d}_F$ are defined in Condition 2(i), and D is defined in (B.26).

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

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

Its derivative is then

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

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

$$\begin{aligned}
J_{1,F} =& \mathbb{E}\left[\text{tr}\left(\frac{\partial}{\partial x}\eta_F(\tilde{\xi}_F+\tilde{d}_F)'E\tilde{V}_F\right)\right] \\
=& \mathbb{E}\left[\frac{-\text{tr}(A_F)}{\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right] - 2\mathbb{E}\left[\frac{\text{tr}[D(\tilde{\xi}_F+\tilde{d}_F)(\tilde{\xi}_F+\tilde{d}_F)'E\tilde{V}_F]}{[\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right] \\
=& \mathbb{E}\left[\frac{-\text{tr}(A_F)}{\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right] + 2\mathbb{E}\left[\frac{-(\tilde{\xi}_F+\tilde{d}_F)'E\tilde{V}_FD(\tilde{\xi}_F+\tilde{d}_F)}{[\text{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right].
\end{aligned} \tag{B.37}$$

By the definitions of A_F, D and E, one has

$$\begin{aligned}
&-(\tilde{\xi}_F+\tilde{d}_F)'E\tilde{V}_F D(\tilde{\xi}_F+\tilde{d}_F)\\
=&(\tilde{\xi}_F+\tilde{d}_F)'[\ -I_k \quad I_k\ ]'\Upsilon(V_{F,SP}-cov_F)\Upsilon[\ -I_k \quad I_k\ ](\tilde{\xi}_F+\tilde{d}_F)\\
\leq&\rho_{\max}[\Upsilon^{1/2}(V_{F,SP}-cov_F)\Upsilon^{1/2}](\tilde{\xi}_F+\tilde{d}_F)'[\ -I_k \quad I_k\ ]'\Upsilon[\ -I_k \quad I_k\ ](\tilde{\xi}_F+\tilde{d}_F)\\
=&\rho_{\max}(A_F)(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F),
\end{aligned} \tag{B.38}$$

where the last equality holds since $\rho_{\max}[\Upsilon^{1/2}(V_{F,SP}-cov_F)\Upsilon^{1/2}] = \rho_{\max}[\Upsilon(V_{F,SP}-cov_F)]$ $= \rho_{\max}(A_F)$. Combining the results in (B.37) and (B.38) gives

$$\begin{aligned}
J_{1,F} \leq&\mathbb{E}\left[\frac{-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right]+2\mathbb{E}\left[\frac{\rho_{\max}(A_F)(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right]\\
=&\mathbb{E}\left[\frac{2\rho_{\max}(A_F)-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right]-\mathbb{E}\left[\frac{2\rho_{\max}(A_F)\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right].
\end{aligned} \tag{B.39}$$

Then by applying some algebraic operations to $J_{2,F}$, one gets

$$\begin{aligned}
J_{2,F} =&\mathbb{E}\left[\frac{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)-\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right]\\
=&\mathbb{E}\left[\frac{1}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right]-\mathbb{E}\left[\frac{\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right].
\end{aligned} \tag{B.40}$$

Combining (B.36), (B.39) and (B.40) gives

$$\begin{aligned}
r(u_{F,d}) \leq& 2\mathrm{tr}(A_F)\mathbb{E}\left[\frac{2\rho_{\max}(A_F)-\mathrm{tr}(A_F)}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right]\\
&-2\mathrm{tr}(A_F)\mathbb{E}\left[\frac{2\rho_{\max}(A_F)\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right]\\
&+[\mathrm{tr}(A_F)]^2\mathbb{E}\left[\frac{1}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right]\\
&-[\mathrm{tr}(A_F)]^2\mathbb{E}\left[\frac{\mathrm{tr}(B_F)}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right]\\
=&\mathbb{E}\left[\frac{\mathrm{tr}(A_F)[4\rho_{\max}(A_F)-\mathrm{tr}(A_F)]}{\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)}\right]-\mathbb{E}\left[\frac{\mathrm{tr}(A_F)\mathrm{tr}(B_F)[4\rho_{\max}(A_F)+\mathrm{tr}(A_F)]}{[\mathrm{tr}(B_F)+(\tilde{\xi}_F+\tilde{d}_F)'D(\tilde{\xi}_F+\tilde{d}_F)]^2}\right].
\end{aligned} \tag{B.41}$$

If $\mathrm{tr}(A_F) > 0$ and $\mathrm{tr}(B_F) > 0$, then $\rho_{\max}(A_F) > 0$, and then the second term in (B.41)

will be negative. If, in addition, $\mathrm{tr}(A_F) \geq 4\rho_{\max}(A_F)$, then the first term in (B.41) will be non-negative. This completes the proof of $\sup_{u_{F,d} \in \mathcal{U}} r(u_{F,d}) \leq 0$.

If, furthermore, $\mathrm{tr}(A_F) > 4\rho_{\max}(A_F)$ for some $F \in \mathcal{F}$, then $r(u_{F,d}) < 0$. This indicates that $\inf_{u_{F,d} \in \mathcal{U}} r(u_{F,d}) < 0$. □

References

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

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

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

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

Aït-Sahalia, Y., P. J. Bickel, and T. M. Stoker (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics 105*(2), 363–412.

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

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, 43–72.

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

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

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

Bajari, P., V. Chernozhukov, H. Hong, and D. Nekipelov (2015). Identification and efficient semiparametric estimation of a dynamic discrete game. Technical report, National Bureau of Economic Research.

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

Baranchik, A. J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Technical report, STANFORD UNIV CALIF.

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

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993). *Efficient and adaptive estimation for semiparametric models*, Volume 4. Johns Hopkins University Press Baltimore.

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

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica: Journal of the Econometric Society*, 1443–1458.

Bierens, H. J. and W. Ploberger (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica: Journal of the Econometric Society*, 1129–1151.

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

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

Buchholz, N., M. Shum, and H. Xu (2020). Semiparametric estimation of dynamic discrete choice models. *Journal of Econometrics*.

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

Chen, X. and Y. Fan (1999). Consistent hypothesis testing in semiparametric and nonparametric models for econometric time series. *Journal of Econometrics 91*(2), 373–401.

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

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

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

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

Claeskens, G. and R. J. Carroll (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika 94*(2), 249–265.

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

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

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

Fan, J., C. Zhang, and J. Zhang (2001). Generalized likelihood ratio statistics and wilks phenomenon. *Annals of statistics*, 153–193.

Fan, Y. and Q. Li (1996, July). Consistent model specification test: Omitted variables and semiparametric functional forms. *Econometrica 64*(4), 865–890.

Fan, Y. and O. Linton (2003). Some higher-order theory for a consistent non-parametric model specification test. *Journal of Statistical Planning and Inference 109*(1-2), 125–154.

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

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

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

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

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

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*.

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

Hart, J. (2013). *Nonparametric smoothing and lack-of-fit tests*. Springer Science & Business Media.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pp. 475–492. NBER.

Heckman, J. J. (1977, March). Sample selection bias as a specification error (with an application to the estimation of labor supply functions). *NBER Working Paper* (172).

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

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

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

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

Hong, Y. and H. White (1995, September). Consistent specification testing via nonparametric series regression. *Econometrica 63*(5), 1133–1159.

Horowitz, J. L. and V. G. Spokoiny (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica 69*(3), 599–631.

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

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

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

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

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 361–379.

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

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

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

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

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

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Lavergne, P. and Q. Vuong (2000). Nonparametric significance testing. *Econometric Theory 16*(4), 576–601.

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

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

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

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

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

Li, Q., C. Hsiao, and J. Zinn (2003). Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics 112*(2), 295–325.

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

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

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

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

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

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

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

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 1349–1382.

Newey, W. K. (1995). Convergence rates & asymptotic normality for series estimators.

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

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

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

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

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

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

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

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

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

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

Ritov, Y. and P. J. Bickel (1990). Achieving information bounds in non and semiparametric models. *The Annals of statistics 18*(2), 925–938.

Robins, J. M. and A. Rotnitzky (2001). Comment on the bickel and kwon article,"inference for semiparametric models: Some questions and an answer". *Statistica Sinica 11*(4), 920–936.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

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

Rosenbaum, P. R. and D. B. Rubin (1983a). 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.

Rosenbaum, P. R. and D. B. Rubin (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

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

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

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University.

Stinchcombe, M. B. and H. White (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric theory 14*(3), 295–325.

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

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Vansteelandt, S., M. Bekaert, and G. Claeskens (2012). On model selection and model misspecification in causal inference. *Statistical methods in medical research 21*(1), 7–30.

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

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

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

Wooldridge, J. M. (1992). A test for functional form against nonparametric alternatives. *Econometric Theory 8*(4), 452–475.

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*, 783–809.

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

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