Quantitative Economics 10 (2019), 931–979 1759-7331/20190931

On uniform asymptotic risk of averaging GMM estimators

Xu Cheng
Department of Economics, University of Pennsylvania

Zhipeng Liao
Department of Economics, UCLA

Ruoyao Shi
Department of Economics, UC Riverside

This paper studies the averaging GMM estimator that combines a conservative GMM estimator based on valid moment conditions and an aggressive GMM estimator based on both valid and possibly misspecified moment conditions, where the weight is the sample analog of an infeasible optimal weight. We establish asymptotic theory on uniform approximation of the upper and lower bounds of the finite-sample truncated risk difference between any two estimators, which is used to compare the averaging GMM estimator and the conservative GMM estimator. Under some sufficient conditions, we show that the asymptotic lower bound of the truncated risk difference between the averaging estimator and the conservative estimator is strictly less than zero, while the asymptotic upper bound is zero uniformly over any degree of misspecification. The results apply to quadratic loss functions. This uniform asymptotic dominance is established in non-Gaussian semiparametric nonlinear models.

Keywords. Asymptotic risk, finite-sample risk, generalized shrinkage estimator, GMM, misspecification, model averaging, nonstandard estimator, uniform approximation.

JEL classification. C13, C36, C52.

1. Introduction

We are interested in estimating some finite dimensional parameter $\theta_F \in \mathbb{R}^{d_\theta}$ which is uniquely identified by the moment restrictions

$$\mathbb{E}_F\big[g_1(W, \theta_F)\big] = 0_{r_1 \times 1} \tag{1.1}$$

Xu Cheng: xucheng@upenn.edu
Zhipeng Liao: zhipeng.liao@econ.ucla.edu
Ruoyao Shi: ruoyao.shi@ucr.edu

We thank Benedikt Pötscher for the insightful comments that greatly improved the quality of the paper. We appreciate useful suggestions from three anonymous referees. The paper also benefited from discussions with Donald Andrews, Denis Chetverikov, Patrik Guggenberger, Jinyong Hahn, Jia Li, Rosa Matzkin, Hyunsik Roger Moon, Ulrich Mueller, Joris Pinkse, Frank Schorfheide, Shuyang Sheng, and participants at various seminars and conferences.

for some known vector function $g_1(\cdot): \mathcal{W} \times \Theta \to \mathbb{R}^{r_1}$, where Θ is a compact subset of $\mathbb{R}^{d_\theta}$, W is a random vector with support $\mathcal{W}$ and joint distribution F, and $\mathbb{E}_F[\cdot]$ denotes the expectation operator under F. Suppose we have i.i.d. data $\{W_i\}_{i=1}^n$, where W_i has distribution F for $i = 1, \ldots, n$.[1] Let $\overline{g}_1(\theta) = n^{-1} \sum_{i=1}^n g_1(W_i, \theta)$. An efficient GMM estimator for θ_F is

$$\widehat{\theta}_1 = \underset{\theta \in \Theta}{\arg\min}\, \overline{g}_1(\theta)'(\overline{\Omega}_1)^{-1}\overline{g}_1(\theta), \tag{1.2}$$

where $\overline{\Omega}_1 = n^{-1} \sum_{i=1}^n g_1(W_i, \widetilde{\theta}_1) g_1(W_i, \widetilde{\theta}_1)' - \overline{g}_1(\widetilde{\theta}_1)\overline{g}_1(\widetilde{\theta}_1)'$ is the efficient weighting matrix with some preliminary consistent estimator $\widetilde{\theta}_1$.[2] In a linear instrumental variable (IV) example, $Y = X'\theta_F + U$ where the IV $Z_1 \in \mathbb{R}^{r_1}$ satisfies $\mathbb{E}_F[Z_1 U] = 0_{r_1 \times 1}$. The moments in (1.1) hold with $g_1(W, \theta) = Z_1(Y - X'\theta)$ and θ_F is uniquely identified if $\mathbb{E}_F[Z_1 X']$ has full column rank. Under certain regularity conditions, it is well known that $\widehat{\theta}_1$ is consistent and achieves the lowest asymptotic variance among GMM estimators based on the moments in (1.1); see Hansen (1982).

If one has additional moments,

$$\mathbb{E}_F\big[g^*(W, \theta_F)\big] = 0_{r^* \times 1} \tag{1.3}$$

for some known function $g^*(\cdot): \mathcal{W} \times \Theta \to \mathbb{R}^{r^*}$, imposing them together with (1.1) can further reduce the asymptotic variance of the GMM estimator. However, if these additional moments are misspecified in the sense that $\mathbb{E}_F[g^*(W, \theta_F)] \neq 0_{r^* \times 1}$, imposing (1.3) may result in inconsistent estimation. The choice of moment conditions is routinely faced by empirical researchers. Take the linear IV model for example. One typically starts with a large number of candidate IVs but only has confidence that a small number of them are valid, denoted by Z_1. The rest of them, denoted by Z^*, are valid only under certain economic hypothesis that yet to be tested. In this example, $g^*(W, \theta) = Z^*(Y - X'\theta)$. In contrast to the conservative estimator $\widehat{\theta}_1$, an aggressive estimator $\widehat{\theta}_2$ always imposes (1.3) regardless of its validity. Let $g_2(W_i, \theta) = (g_1(W_i, \theta)', g^*(W_i, \theta)')'$ for $i = 1, \ldots, n$, and $\overline{g}_2(\theta) = n^{-1} \sum_{i=1}^n g_2(W_i, \theta)$. The aggressive estimator $\widehat{\theta}_2$ takes the form

$$\widehat{\theta}_2 = \underset{\theta \in \Theta}{\arg\min}\, \overline{g}_2(\theta)'(\overline{\Omega}_2)^{-1}\overline{g}_2(\theta), \tag{1.4}$$

where $\overline{\Omega}_2$ is constructed in the same way as $\overline{\Omega}_1$ except that $g_1(W_i, \theta)$ is replaced by $g_2(W_i, \theta)$.[3]

Because imposing $\mathbb{E}_F[g^*(W, \theta_F)] = 0_{r^* \times 1}$ is a double-edged sword, a data-dependent decision usually is made to choose between $\widehat{\theta}_1$ and $\widehat{\theta}_2$. To study such a decision and the subsequent estimator, let

$$\delta_F = \mathbb{E}_F\big[g^*(W, \theta_F)\big] \in \mathbb{R}^{r^*}. \tag{1.5}$$

[1] The main theory of the paper can be easily extended to time series models with dependent data, as long as the preliminary results in Lemma B.1 hold.

[2] For example, $\widetilde{\theta}_1$ could be the GMM estimator similar to $\widehat{\theta}_1$ but with an identity weighting matrix; see (B.5) in the Appendix.

[3] See the first line of equations (E.13) in the Supplemental Appendix for the definition of $\overline{\Omega}_2$. In particular, $\overline{\Omega}_2$ is constructed using $\widetilde{\theta}_1$, the preliminary consistent estimator based on the valid moment conditions in (1.1) only.

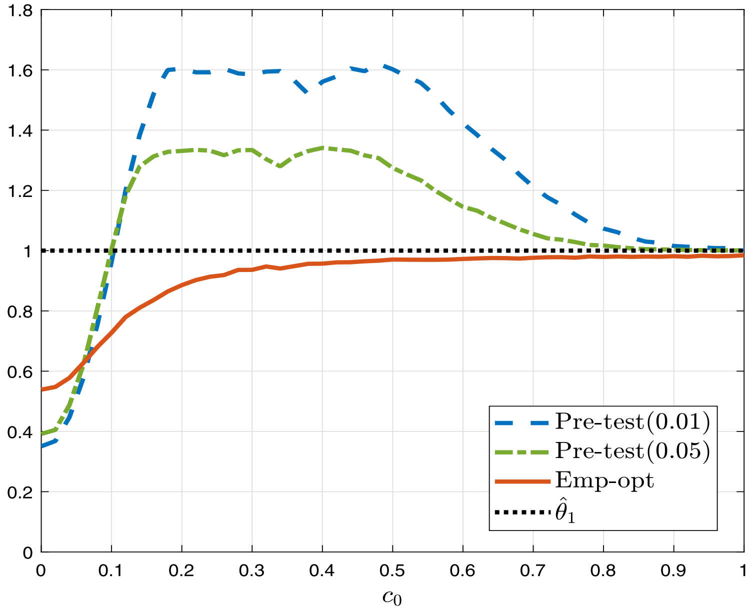

FIGURE 1. Finite sample ($n = 500$) MSEs of the pre-test and the averaging GMM estimators. *Note*: "Pre-test(0.01)" and "Pre-test(0.05)" refer to the pre-test GMM estimator based on the J-test statistic $n\overline{g}_2(\widehat{\theta}_2)'(\overline{\Omega}_2)^{-1}\overline{g}_2(\widehat{\theta}_2)$ with nominal size 0.01 and 0.05, respectively. "Emp-opt" refers to the averaging GMM estimator with weight defined in (4.3) of the paper. In this simulation, we set $\delta_F = c_0\omega$ where c_0 is in [0,1] and ω is a real vector. At each c_0, we consider 127 different values for ω and report the largest finite sample MSEs of the estimators. Details of the simulation design for this figure is provided in Section 6.1.

The pre-testing approach tests the null hypothesis $H_0 : \delta_F = 0_{r^*\times 1}$ and constructs an estimator

$$\widehat{\theta}_{\text{pre}} = 1\{T_n > c_\alpha\}\widehat{\theta}_1 + 1\{T_n \le c_\alpha\}\widehat{\theta}_2 \tag{1.6}$$

for some test statistic T_n with the critical value c_α at the significance level α. One popular test is the J-test (see Hansen (1982)), and c_α is the $1-\alpha$ quantile of the chi-squared distribution with degree of freedom $r_2 - d_\theta$ where $r_2 = r_1 + r^*$. Because the power of this test against the fixed alternative is 1, $\widehat{\theta}_{\text{pre}}$ equals $\widehat{\theta}_1$ with probability 1 asymptotically ($n \to \infty$) for those fixed misspecified model where $\delta_F \ne 0_{r^*\times 1}$. Thus, it seems that $\widehat{\theta}_{\text{pre}}$ is immune to moment misspecification. However, we care about the finite-sample mean squared error (MSE) of $\widehat{\theta}_{\text{pre}}$ in practice and this standard pointwise asymptotic analysis (δ_F is fixed and $n \to \infty$) provides a poor approximation to the former.[4] To see the comparison between $\widehat{\theta}_{\text{pre}}$ and $\widehat{\theta}_1$, the dashed line and the dashed-dotted line in Figure 1 plot the maximum finite-sample ($n = 500$) MSEs of $\widehat{\theta}_{\text{pre}}$ with $\alpha = 0.01$ and 0.05, respectively,

[4] The poor approximation of the pointwise asymptotics to the finite sample properties of the pre-test estimator has been noted in Shibata (1986), Pötscher (1991), Kabaila (1995, 1998), and Leeb and Pötscher (2005, 2008), among others.

while the MSE of $\widehat{\theta}_1$ is normalized to 1.[5] For some values of δ_F, the MSE of $\widehat{\theta}_{\text{pre}}$ may be larger than that of $\widehat{\theta}_1$, sometimes by more than 50%. Note that the pre-test estimators exhibit multiple peaks because the simulation design allows for multiple potentially misspecified moments and considers two different ways of parametrizing δ_F. Given c_0, the norm of δ_F may be different in the two different parametrizations.

The goal of this paper is twofold. First, we propose a data-dependent averaging of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ that takes the form

$$\widehat{\theta}_{\text{eo}} = (1 - \widetilde{\omega}_{\text{eo}})\widehat{\theta}_1 + \widetilde{\omega}_{\text{eo}}\widehat{\theta}_2, \tag{1.7}$$

where $\widetilde{\omega}_{\text{eo}} \in [0, 1]$ is a data-dependent weight specified in (4.7) below. The subscript in $\widetilde{\omega}_{\text{eo}}$ is short for empirical optimal because this weight is an empirical analog of an infeasible optimal weight ω_F^* defined in (4.3) below. We plot the finite-sample MSE of this averaging estimator as the solid line in Figure 1. This averaging estimator is robust to misspecification in the sense that the solid line is below 1 for all values of δ_F, in contrast to the bump in the dashed line that represents the pre-test estimator. Second, we develop a *uniform* asymptotic theory to justify the finite-sample robustness of this averaging estimator. We quantify the upper and lower bounds of the asymptotic risk differences between the averaging estimator and the conservative estimator, and show that this averaging estimator dominates the conservative estimator uniformly over a large class of models with different degrees of misspecification in certain asymptotic sense.[6] The uniform dominance is established under the truncated weighted loss function which is defined in (3.11) below.[7] Our uniform dominance result relies on the asymptotic properties of the GMM estimators, therefore, it is weaker than the exact finite sample dominance result of the James–Stein estimator established in the Gaussian sampling models.

The rest of the paper is organized as follows. Section 2 discusses the literature related to our paper. Section 3 defines the parameter space over which the uniform result is established and defines uniform dominance. Section 4 introduces the averaging weight. Section 5 provides an analytical representation of the bounds of the asymptotic risk differences and applies it to show that the averaging GMM estimator uniformly dominates the conservative estimator. Section 6 investigates the finite sample performance of our averaging estimator using Monte Carlo simulations. Section 7 concludes. Proof of the main results of the paper and additional simulation results are given in the Appendix. Analysis of the pre-test estimator, extra simulation studies, and proofs of some auxiliary results are included in the Online Supplemental Material of the paper (Cheng, Liao, and Shi (2019)).

Notation. For any real matrix A, we use $\|A\|$ to denote the Frobenius norm of A, that is, $\|A\| = (\text{tr}(A'A))^{1/2}$ where $\text{tr}(\cdot)$ denotes the trace operator of square matrices. If A is a real symmetric matrix, $\rho_{\min}(A)$ and $\rho_{\max}(A)$ denote the smallest and largest

[5] That is, the dashed line and the dashed-dotted line represent the ratios of the maximum MSEs of the two pre-test estimators divided by the MSE of $\widehat{\theta}_1$, respectively.

[6] The lower and upper bounds of asymptotic risk difference are defined in (3.12) below.

[7] Truncation at a large number is needed for the asymptotic analysis of the risk of general estimator without imposing stringent conditions such as uniform integrability.

eigenvalues of A, respectively. For any positive integers d_1 and d_2, I_{d_1} and $0_{d_1 \times d_2}$ stand for the $d_1 \times d_1$ identity matrix and $d_1 \times d_2$ zero matrix, respectively. Let $\mathrm{vec}(\cdot)$ denote vectorization of a matrix and $\mathrm{vech}(\cdot)$ denotes the half- vectorization of a symmetric matrix. Let $\mathbb{R} = (-\infty, +\infty)$, $\mathbb{R}_+ = [0, +\infty)$, $\mathbb{R}_\infty = \mathbb{R} \cup \{\pm\infty\}$, and $\mathbb{R}_{+,\infty} = \mathbb{R}_+ \cup \{+\infty\}$. For any positive integer d and any set $\mathbb{S}$, $\mathbb{S}^d$ denotes the Cartesian product of d many sets: $\mathbb{S}_1 \times \cdots \times \mathbb{S}_d$ with $\mathbb{S}_j = \mathbb{S}$ for $j = 1, \ldots, d$. For any finite positive integer d and any set $\mathbb{S} \subset \mathbb{R}^d$, $\mathrm{int}(\mathbb{S})$ denotes the interior of $\mathbb{S}$ under the Euclidean norm. We use $\mathbb{N}$ to denote the set of natural numbers and $\{p_n\} = \{p_n : n \in \mathbb{N}\}$ denote a subsequence of $\{n\}_{n \in \mathbb{N}}$. For any (possibly random) positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, $a_n = O_p(b_n)$ means that $\sup_{n \in \mathbb{N}} \Pr(a_n/b_n > c) \to 0$ as $c \to \infty$; $a_n = o_p(b_n)$ means that for all $\varepsilon > 0$, $\lim_{n \to \infty} \Pr(a_n/b_n > \varepsilon) = 0$. Let "$\to_p$" and "$\to_D$" stand for convergence in probability and convergence in distribution, respectively. The notation $a \equiv b$ means a is defined as b.

2. Related literature

Our uniform dominance result is related to the Stein's phenomenon (Stein (1956)) in parametric models. The James–Stein (JS) estimator shrinks the maximum likelihood estimator (MLE) toward zero and has been shown to dominate the MLE in exact normal sampling; see James and Stein (1961). Green and Strawderman (1991) proposed an averaging estimator in the Gaussian location model which shrinks an unbiased estimator toward a biased estimator with a JS type of weight, and showed that the averaging estimator dominates the unbiased estimator. Green and Strawderman (1991) assumed that the unbiased estimator is independent of the biased estimator, which is relaxed in Kim and White (2001), Judge and Mittelhammer (2004), and Mittelhammer and Judge (2005).[8, 9] These papers propose averaging estimators which shrink asymptotically unbiased estimators toward biased estimators in the semiparametric setting, and show that the averaging estimators based on infeasible weights dominate the unbiased estimators in the Gaussian location models. Kim and White (2001) showed that the infeasible weight can be consistently estimated when the asymptotic bias of the biased estimator is zero. Judge and Mittelhammer (2004) and Mittelhammer and Judge (2005) provided approximators of the infeasible optimal weights and show that these approximators can be consistently estimated. These estimators and our estimator are all linear combinations of the unbiased estimator and the biased estimator. However, even in the Gaussian location models, the weights are different and their sufficient conditions and proofs for dominance are different; see Appendix A for details.[10]

Hansen (2016) considered the JS-type averaging estimator in general parametric models and shows the Stein-dominance result in a pointwise local asymptotic

[8] We thank an anonymous referee who referred Green and Strawderman (1991) and Judge and Mittelhammer (2004) to us.

[9] Judge and Mittelhammer (2007) proposed an averaging estimator which combine different GMM estimators with weights determined by the empirical likelihood method. However, the properties of this averaging estimator are not fully investigated and no dominace results are established in this paper.

[10] In the Gaussian location model, our dominance results require $d_\theta \geq 4$; Green and Stawderman (1991) required $d_\theta \geq 3$ by imposing independence between the unbiased and the biased estimators; and Kim and White (2001), Judge and Mittelhammer (2004) and Mittelhammer and Judge (2005) required $d_\theta \geq 5$.

sense.[11] Hansen (2017) proposed an averaging estimator that combines the ordinary least squares (OLS) estimator and the two-stage-least-squares (2SLS) estimator in linear IV models, and shows that the averaging estimator has smaller local asymptotic risk than the OLS estimator. DiTraglia (2016) also studied the averaging GMM estimator in the pointwise local asymptotic framework. The averaging weight of his estimator is based on the focused moment selection criterion with a targeted parameter. The simulation results in the paper show that this averaging estimator does not uniformly dominate the conservative estimator. Many other frequentist model averaging estimators are studied in the literature, including Buckland, Burnham, and Augustin (1997), Hjort and Claeskens (2003, 2006), Hansen (2007), Claeskens and Carroll (2007), Hansen and Racine (2012), Cheng and Hansen (2015), Lu and Su (2015), to name only a few.

Different from the aforementioned papers, our paper is the first to show global dominance based on uniform asymptotic approximation.[12] This uniform analysis is similar to those studied in Andrews, Cheng, and Guggenberger (2011) for uniform size control for inference in nonstandard problems, but the present paper is for estimation rather than inference and focuses on a misspecification issue that is not studied in these papers.

The uniform dominance property of the averaging estimator does not contradict the risk properties of the post-model-selection estimators found in Yang (2005) and Leeb and Pötscher (2008). Measured by the MSE, the post-model-selection estimator usually does better than the unrestricted estimator in part of the parameter space and worse than the latter in other part of the parameter space. One standard example is the Hodge's estimator, whose scaled maximal MSE diverges to infinity with the growth of the sample size (see, e.g., Lehmann and Casella (1998)). Similar unbounded risk results are established in Yang (2005) and Leeb and Pötscher (2008) for post-model-selection estimator based on consistent model selection procedures. Such estimators have unbounded (scaled) maximal MSE because given the consistent model selection procedure: (i) there exist DGPs where the restrictions to be tested/selected are (locally) misspecified; (ii) the model selection procedures select these misspecified restrictions with high probabilities, converging to 1 asymptotically; (iii) the restricted estimator has unbounded (scaled) MSE under these DGPs.[13] In contrast, the empirical optimal weight of our averaging

[11] For a given real vector d, the pointwise local asymptotic analysis considers a sequence of local DGPs $\{F_n\}_n$ under which $\delta_{F_n} = dn^{-1/2}$, and derives the asymptotic (truncated) risk of the averaging estimator under $\{F_n\}_n$ for the given d. Such analysis will produce a pointwise risk function (on d) for the averaging estimator. Evaluation of the averaging estimator is then conducted using the pointwise local asymptotic risk function.

[12] In the uniform global asymptotic framework, one has to study the asymptotic behavior of the supermum and the infimum of the finite sample risk of the averaging estimator, where the supermum and the infimum are taken over a class of DGPs which include both the locally misspecified and many more severely misspecified DGPs. See Section 3 for more details.

[13] The post-model-selection estimator based on a conservative model selection procedure (e.g., hypothesis test with fixed critical value or Akaike information criterion) typically do not have unbounded (scaled) maximal MSE. However, its asymptotic maximal MSE is not guaranteed to be less than or equal to the benchmark estimator (e.g., the conservative GMM estimator in the framework of this paper). The pre-test estimators in Figure 1 are good examples, since they are based on the J-test with nominal size 0.01 and 0.05.

estimator is based on an infeasible optimal weight that satisfies: (i) when the aggressive/restricted GMM estimator has unbounded (scaled) MSE, the averaging weight on it is small, converging to 0 asymptotically. The resulting averaging estimator has the same asymptotic properties as the conservative GMM estimator; (ii) the Stein's dominance result applies in the asymptotic sense. Hence our averaging estimator is essentially different from the post-model-selection estimator.

There is a large literature studying the validity of GMM moment conditions. Many methods can be applied to detect the validity, including the overidentification tests (see, e.g., Sargan (1958), Hansen (1982), and Eichenbaum, Hansen, and Singleton (1988)), the information criteria (see, e.g., Andrews (1999), Andrews and Lu (2001), and Hong, Preston, and Shum (2003)), and the penalized estimation methods (see, e.g., Liao (2013), Cheng and Liao (2015)). Recently, misspecified moments and their consequences are considered by Ashley (2009), Berkowitz, Caner, and Fang (2012), Conley, Hansen, and Rossi (2012), Doko Tchatoka and Dufour (2008, 2014), Guggenberger (2012), Nevo and Rosen (2012), Kolesar, Chetty, Friedman, Glaeser, Imbens (2015), Small (2007), Small, Cai, Zhang, and Kang (2016), among others. Moon and Schorfheide (2009) explored overidentifying moment inequalities to reduce the MSE. This paper contributes to this literature by providing new uniform results for potentially misspecified semiparametric models.

There is also a large literature studying adaptive estimation in nonparametric regression model using model averaging; see Yang (2000, 2003, 2004), Leung and Barron (2006), and the references therein. Since the unknown function can be written as a linear combination of (possibly infinitely but countably many) basis functions, the nonparametric model may be well approximated by parametric regression models in finite samples. These papers show that the averaging estimators which combine OLS estimators from different parametric models with data dependent weights may achieve the optimal convergence rate up to some logarithm factor. Our paper is different from these papers since the parameter of interest in our paper is a finite dimensional real value, not an unknown function, and the bias and variance trade-off of our averaging estimator is due to the possibly misspecified moment conditions. Moreover, there is a benchmark estimator in our paper, that is, the conservative GMM estimator whose asymptotic properties are well known. Our goal is to propose an averaging estimator with uniformly smaller risk than the conservative estimator.

3. Parameter space and uniform dominance

Let $g_{2,j}(w,\theta)$ $(j=1,\dots,r_2)$ denote the jth component function of $g_2(w,\theta)$. We assume that $g_{2,j}(w,\theta)$ for $j=1,\dots,r_2$ is twice continuously differentiable with respect to θ for any $w\in\mathcal{W}$. The first- and second-order derivatives of $g_2(w,\theta)$ with respect to θ are denoted by

$$g_{2,\theta}(w,\theta)\equiv\begin{pmatrix}\dfrac{\partial g_{2,1}(w,\theta)}{\partial\theta'}\\ \vdots\\ \dfrac{\partial g_{2,r_2}(w,\theta)}{\partial\theta'}\end{pmatrix}\quad\text{and}\quad g_{2,\theta\theta}(w,\theta)\equiv\begin{pmatrix}\dfrac{\partial^2 g_{2,1}(w,\theta)}{\partial\theta\,\partial\theta'}\\ \vdots\\ \dfrac{\partial^2 g_{2,r_2}(w,\theta)}{\partial\theta\,\partial\theta'}\end{pmatrix},\tag{3.1}$$

respectively.[14] Let $\mathcal{F}$ be a set of distribution functions on $\mathcal{W}$. For $k = 1$ and 2, define the expectation of the moment functions, the Jacobian matrix and the variance–covariance matrix as

$$\begin{aligned} M_{k,F} &\equiv \mathbb{E}_F\big[g_k(W, \theta_F)\big], \\ G_{k,F} &\equiv \mathbb{E}_F\big[g_{k,\theta}(W, \theta_F)\big], \quad \text{and} \\ \Omega_{k,F} &\equiv \mathbb{E}_F\big[g_k(W, \theta_F)g_k(W, \theta_F)'\big] - M_{k,F}M'_{k,F} \end{aligned} \tag{3.2}$$

for any $F \in \mathcal{F}$, respectively. The moments above exist by Assumption 3.2 below.

Let

$$Q_F(\theta) \equiv \mathbb{E}_F\big[g_2(W, \theta)\big]'\Omega_{2,F}^{-1}\mathbb{E}_F\big[g_2(W, \theta)\big] \tag{3.3}$$

for any $\theta \in \Theta$, which denotes the population criterion of the GMM estimation in (1.4). For any $\theta \in \Theta$, define $B_\varepsilon^c(\theta) = \{\theta^* \in \Theta : \|\theta^* - \theta\| \geq \varepsilon\}$. We consider the risk difference between two estimators uniformly over $F \in \mathcal{F}$, where $\mathcal{F}$ satisfies Assumptions 3.1–3.3 below.

Assumption 3.1. *The following conditions hold*:

(i) *for any* $F \in \mathcal{F}$, $\mathbb{E}_F[g_1(W, \theta_F)] = 0_{r_1\times 1}$ *for some* $\theta_F \in \text{int}(\Theta)$;

(ii) *for any* $\varepsilon > 0$, $\inf_{F\in\mathcal{F}} \inf_{\theta\in B_\varepsilon^c(\theta_F)} \|\mathbb{E}_F[g_1(W, \theta)]\| > 0$;

(iii) *for any* $F \in \mathcal{F}$, *there is* $\theta_F^* \in \text{int}(\Theta)$ *such that*

$$\inf_{F\in\mathcal{F}} \inf_{\theta\in B_\varepsilon^c(\theta_F^*)} \big[Q_F(\theta) - Q_F\big(\theta_F^*\big)\big] > 0 \quad \textit{for any } \varepsilon > 0;$$

(iv) $\inf_{\{F\in\mathcal{F}:\|\delta_F\|>0\}} \frac{\|G'_{2,F}\Omega_{2,F}^{-1}\delta_{2,F}\|}{\|\delta_{2,F}\|^\tau} > 0$ *where* $\delta'_{2,F} = (0_{1\times r_1}, \delta'_F)$ *and* $\tau > 0$ *is a fixed constant*;

(v) $0_{r^*\times 1} \in \text{int}(\Delta_\delta)$ *where* $\Delta_\delta = \{\delta_F : F \in \mathcal{F}\}$.

Assumptions 3.1(i)–(ii) require that the true unknown parameter θ_F is uniquely identified by the moment conditions $\mathbb{E}_F[g_1(W, \theta_F)] = 0_{r_1\times 1}$ for any DGP $F \in \mathcal{F}$. Assumption 3.1(iii) implies that for any $F \in \mathcal{F}$, a pseudo-true value θ_F^* is identified by the unique minimizer of the population GMM criterion $Q_F(\theta)$ under possible misspecification. Assumption 3.1(iv) requires that $\delta_{2,F}$ is not orthogonal to $\Omega_{2,F}^{-1}G_{2,F}$, which rules out the special case that θ_F may be consistently estimable even with severely misspecified moment conditions. Assumption 3.1(v) implies that the set of distribution functions $\mathcal{F}$ is rich such that it includes the distributions under which the extra moment conditions are correctly specified, locally misspecified or severely misspecified. Uniform dominance can be easily established if we only allow for correctly specified models or severely misspecified models, because the desired dominance results hold trivially following a pointwise analysis. Assumption 3.1(v) ensures that the extra moment conditions may have different degrees of misspecification in the parameter space.

[14] By definition, $g_{1,\theta}(w, \theta)$ and $g_{1,\theta\theta}(w, \theta)$ are the leading $r_1 \times d_\theta$ and $(r_1 d_\theta) \times d_\theta$ submatrices of $g_{2,\theta}(w, \theta)$ and $g_{2,\theta\theta}(w, \theta)$, respectively.

Assumption 3.2. *The following conditions hold*:

(i) *For* $j = 1, \dots, r_2$, $g_{2,j}(w, \theta)$ *is twice continuously differentiable with respect to* θ *for any* $w \in \mathcal{W}$;

(ii) $\sup_{F\in\mathcal{F}} \mathbb{E}_F[\sup_{\theta\in\Theta}(\|g_2(W, \theta)\|^{2+\gamma} + \|g_{2,\theta}(W, \theta)\|^{2+\gamma} + \|g_{2,\theta\theta}(W, \theta)\|^{2+\gamma})] < \infty$ *for some* $\gamma > 0$;

(iii) $\inf_{F\in\mathcal{F}} \rho_{\min}(\Omega_{2,F}) > 0$;

(iv) $\inf_{F\in\mathcal{F}} \rho_{\min}(G'_{1,F}G_{1,F}) > 0$.

Assumption 3.2(i) requires that the moment functions are smooth. Assumption 3.2(ii) imposes $2 + \gamma$ finite moment conditions on the GMM moment functions and their first and second derivatives. Assumptions 3.2(iii) and 3.2(iv) are important sufficient conditions for the local identification of the unknown parameter in GMM with valid moment conditions.

The next assumption is on the nuisance parameters of the DGP $F \in \mathcal{F}$. Write

$$v_F = \left(\text{vec}(G_{2,F})', \text{vech}(\Omega_{2,F})', \delta'_F\right) \tag{3.4}$$

for any $F \in \mathcal{F}$. It is clear that v_F includes the Jacobian matrix, the variance–covariance matrix, and the measure of misspecification of the moment conditions $\mathbb{E}_F[g^*(W, \theta_F)] = 0_{r^*\times 1}$. Let

$$\overline{v}_F = \left(\text{vec}(G_{2,F})', \text{vech}(\Omega_{2,F})'\right) \tag{3.5}$$

for any $F \in \mathcal{F}$.

Assumption 3.3. *The following conditions hold*:

(i) *For any* $F \in \mathcal{F}$ *with* $\delta_F = 0_{r^*\times 1}$, *there exists a constant* $\varepsilon_F > 0$ *such that for any* $\widetilde{\delta} \in \mathbb{R}^{r^*}$ *with* $0 \leq \|\widetilde{\delta}\| < \varepsilon_F$, *there is* $\widetilde{F} \in \mathcal{F}$ *with* $\delta_{\widetilde{F}} = \widetilde{\delta}$ *and* $\|\overline{v}_{\widetilde{F}} - \overline{v}_F\| \leq C\|\widetilde{\delta}\|^\kappa$ *for some* $\kappa > 0$;

(ii) *The set* $\Lambda \equiv \{v_F: F \in \mathcal{F}\}$ *is closed.*

Assumption 3.3(i) requires that for any $F \in \mathcal{F}$ such that $\mathbb{E}_F[g_2(W, \theta_F)] = 0_{r^*\times 1}$ is valid, there are many DGPs $\widetilde{F} \in \mathcal{F}$ which are close to F. Here, the closeness of any two DGPs F and $\widetilde{F}$ is measured by the distance between v_F and $v_{\widetilde{F}}$. Assumption 3.3(i) and (ii) are useful to derive the exact expression of the asymptotic risk of the GMM estimator.

Example 3.1 (Linear IV model). We study a linear IV model and provide a set of low-level conditions that imply Assumptions 3.1, 3.2, and 3.3. The parameters of interest θ_0 are the coefficients of the endogenous regressors X in

$$Y = X'\theta_0 + U, \tag{3.6}$$

with some valid IVs $Z_1 \in \mathbb{R}^{r_1}$ and some potentially misspecified IVs $Z^* \in \mathbb{R}^{r^*}$ such that

$$\mathbb{E}_{F^*}[U] = 0, \qquad \mathbb{E}_{F^*}[Z_1U] = 0_{r_1\times 1} \quad \text{and} \tag{3.7}$$

$$Z^* = U\delta_0 + V, \quad \text{with } \mathbb{E}_{F^*}[V] = 0_{r^*\times 1} \text{ and } \mathbb{E}_{F^*}[VU] = 0_{r^*\times 1}, \tag{3.8}$$

where F^* denotes the joint distribution of $(X', Z_1', V', U)'$. In the reduced-form equation (3.8), δ_0 is a $r^* \times 1$ real vector which characterizes the degree of misspecification. Let $\mathcal{F}^*$ denote a class of distributions containing F^*, and let Θ and Δ_δ denote the parameter spaces of θ_0 and δ_0, respectively. The joint distribution of $W = (Y, Z_1', Z^{*\prime}, X')'$ is denoted as F which is determined by θ_0, δ_0, and F^* through the linear equations in (3.6) and (3.8).

For ease of discussion, we further assume that the random vector $(X', Z_1', V', U)'$ follows the normal distribution with mean ϕ and variance–covariance matrix Ψ. Under the normality assumption, each distribution F^* corresponds to a pair of ϕ and Ψ.

For notational simplicity, in Lemma 3.1 below, for any finite dimensional random vectors a_1 and a_2, let $\phi_{a_j} = \mathbb{E}_{F^*}[a_j]$ for $j = 1, 2$, $\Gamma_{a_1 a_2} = \mathbb{E}_{F^*}[a_1 a_2']$, and $\Omega_{a_1 a_2} = \mathbb{E}_{F^*}[a_1 a_2'] - \phi_{a_1}\phi_{a_2}'$.

Lemma 3.1. *Let $\mathcal{F}^*$ denote the set of normal distributions which satisfies*:

(i) $\phi_u = 0$, $\Gamma_{z_1 u} = 0_{r_1 \times 1}$ *and* $\Gamma_{vu} = 0_{r^* \times 1}$;

(ii) $\inf_{F^* \in \mathcal{F}^*} \rho_{\min}(\Gamma_{x z_1}\Gamma_{z_1 x}) > 0$, $\sup_{F^* \in \mathcal{F}^*} \|\phi\|^2 < \infty$ *and* $0 < \inf_{F^* \in \mathcal{F}^*} \rho_{\min}(\Psi) \le \sup_{F^* \in \mathcal{F}^*} \rho_{\max}(\Psi) < \infty$;

(iii) $\inf_{F^* \in \mathcal{F}^*} \inf_{\{\|\delta\| \ge \varepsilon\}} \|\delta\|^{-1} \|(\Gamma_{x z_1}\Gamma_{z_1 z_1}^{-1}\Gamma_{z_1 v} - \Gamma_{xv})\delta - \Gamma_{xu}\| > 0$ *for some $\varepsilon > 0$ that is small enough*;[15]

(iv) $\theta_0 \in \operatorname{int}(\Theta)$ *and Θ is compact and large enough such that the pseudo-true value* $\theta_F^* \in \operatorname{int}(\Theta)$;[16]

(v) $\Delta_\delta = [c_{1,\Delta}, C_{1,\Delta}] \times \cdots \times [c_{r^*,\Delta}, C_{r^*,\Delta}]$ *where $\{c_{j,\Delta}, C_{j,\Delta}\}_{j=1}^{r^*}$ is a set of finite constants with $c_{j,\Delta} < 0 < C_{j,\Delta}$ for $j = 1, \ldots, r^*$*,

then Assumptions 3.1, 3.2, and 3.3 hold.

Condition (i) lists the moment conditions in (3.7) and (3.8). The inequality in Condition (ii) rules out DGPs under which $\rho_{\min}(\Gamma_{x z_1}\Gamma_{z_1 x})$ may be close to zero and (part of) the unknown parameter θ_0 is weakly identified. Condition (ii) also requires that the mean of the random vector $(X', Z_1', V', U)'$ is uniformly bounded and the eigenvalues of its variance–covariance matrix are uniformly finite and bounded away from 0. Condition (iii) requires that the projection residual of the vector Γ_{xu} on the subspace spanned by the matrix $\Gamma_{x z_1}\Gamma_{z_1 z_1}^{-1}\Gamma_{z_1 v} - \Gamma_{xv}$ is bounded away from zero. It is a sufficient condition for Assumption 3.1(iv), which ensures that the aggressive estimator is inconsistent under severe misspecification. Condition (iv) is needed to derive the limit of the aggressive estimator under misspecification. The compactness assumption of Θ is not needed for the linear IV model. However, it is useful to verify Assumptions 3.1, 3.2, and 3.3 which do not

[15] The constant ε depends on the infimum and supremum in Condition (ii) and it is given in (B.3) in the Appendix.

[16] Specific restrictions on Θ which ensure that $\theta_F^* \in \operatorname{int}(\Theta)$ are given in (D.8) and Assumption D.1(vi) in the Supplemental Appendix.

assume any special structure on the model. Condition (v) specifies that the parameter space of δ_0 is a product space.

Lemma 3.1 provides simple conditions on θ_0, δ_0, and $\mathcal{F}^*$ on which uniformity results are subsequently established.[17]

Now we get back to the general set up. For a generic estimator $\widehat{\theta}$ of θ, consider a weighted quadratic loss function

$$\ell(\widehat{\theta}, \theta) = n(\widehat{\theta} - \theta)' \Upsilon (\widehat{\theta} - \theta), \tag{3.9}$$

where Υ is a $d_\theta \times d_\theta$ pre-determined positive semidefinite matrix. For example, if $\Upsilon = I_{d_\theta}$, $\mathbb{E}_F[\ell(\widehat{\theta}, \theta_F)]$ is the MSE of $\widehat{\theta}$. If $\Upsilon = (\Sigma_{1,F} - \Sigma_{2,F})^{-1}$ where $\Sigma_{k,F}$ $(k = 1, 2)$ is defined in (4.4), the weighting matrix Υ rescales $\widehat{\theta}$ by the scale of variance reduction due to the additional moments. If $\Upsilon = \mathbb{E}_F[X_i X_i']$ for regressors X_i, $\mathbb{E}_F[\ell(\widehat{\theta}, \theta_F)]$ is the MSE of $X_i'\widehat{\theta}$, an estimator of $X_i'\theta$.

Below we compare the averaging estimator $\widehat{\theta}_{\text{eo}}$ and the conservative estimator $\widehat{\theta}_1$. We are interested in the bounds of the truncated finite sample risk difference

$$\begin{aligned} \underline{RD}_n(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1; \zeta) &\equiv \inf_{F \in \mathcal{F}} \mathbb{E}_F\big[\ell_\zeta(\widehat{\theta}_{\text{eo}}, \theta_F) - \ell_\zeta(\widehat{\theta}_1, \theta_F)\big] \quad \text{and} \\ \overline{RD}_n(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1; \zeta) &\equiv \sup_{F \in \mathcal{F}} \mathbb{E}_F\big[\ell_\zeta(\widehat{\theta}_{\text{eo}}, \theta_F) - \ell_\zeta(\widehat{\theta}_1, \theta_F)\big], \end{aligned} \tag{3.10}$$

where

$$\ell_\zeta(\widehat{\theta}, \theta_F) \equiv \min\big\{\ell(\widehat{\theta}, \theta_F), \zeta\big\} \tag{3.11}$$

denotes the truncated loss function with an arbitrary trimming parameter ζ. The truncated loss function is employed to facilitate the asymptotic analysis of the bounds of the risk difference. The finite-sample bounds in (3.10) are approximated by

$$\begin{aligned} \text{Asy}\underline{RD}(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1) &\equiv \liminf_{\zeta \to \infty} \liminf_{n \to \infty} \underline{RD}_n(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1; \zeta) \quad \text{and} \\ \text{Asy}\overline{RD}(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1) &\equiv \limsup_{\zeta \to \infty} \limsup_{n \to \infty} \overline{RD}_n(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1; \zeta), \end{aligned} \tag{3.12}$$

which are called lower and upper bounds of the asymptotic risk difference respectively in this paper. The averaging estimator $\widehat{\theta}_{\text{eo}}$ asymptotically uniformly dominates the conservative estimator $\widehat{\theta}_1$ if

$$\text{Asy}\underline{RD}(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1) < 0 \quad \text{and} \quad \text{Asy}\overline{RD}(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1) \le 0. \tag{3.13}$$

The bounds of the asymptotic risk difference build the uniformity over $F \in \mathcal{F}$ into the definition by taking $\inf_{F \in \mathcal{F}}$ and $\sup_{F \in \mathcal{F}}$ before $\liminf_{n\to\infty}$ and $\limsup_{n\to\infty}$, respectively. Uniformity is crucial for the asymptotic results to give a good approximation to their finite-sample counterparts. These uniform bounds are different from pointwise results which are obtained under a fixed DGP. The sequence of DGPs $\{F_n\}$ along which the

[17] Similar results have been established in Section D of the Supplemental Appendix for the linear IV model when the normality assumption on $(X', Z_1', V', U)'$ is relaxed. Section D of the Supplemental Appendix also provides proof for Lemma 3.1 with and without the normality assumption.

supremum or the infimum are approached often varies with the sample size.[18] Therefore, to determine the bounds of the asymptotic risk difference, one has to derive the asymptotic distributions of these estimators under various sequences $\{F_n\}$. Under $\{F_n\}$, the observations $\{W_{n,i}\}_{i=1}^n$ form a triangular array. For notational simplicity, $W_{n,i}$ is abbreviated to W_i throughout the paper.

To study the bounds of asymptotic risk difference, we consider sequences of DGPs $\{F_n\}$ such that δ_{F_n} satisfies[19]

$$\text{(i)} \quad n^{1/2}\delta_{F_n} \to d \in \mathbb{R}^{r^*} \quad \text{or} \quad \text{(ii)} \quad \|n^{1/2}\delta_{F_n}\| \to \infty. \tag{3.14}$$

Case (i) models mild misspecification, where δ_{F_n} is a $n^{-1/2}$-local deviation from $0_{r^* \times 1}$. Case (ii) includes the severe misspecification where $\|\delta_{F_n}\|$ is bounded away from 0 as well as the intermediate case in which $\delta_{F_n} \to 0$ and $\|n^{1/2}\delta_{F_n}\| \to \infty$. To obtain a uniform approximation, all of these sequences are necessary. Once we study the bounds of asymptotic risk difference along each of these sequences, we show that we can glue them together to obtain the bounds of asymptotic risk difference.

4. Averaging weight

We start by deriving the joint asymptotic distribution of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ under different degrees of misspecification. We consider sequences of DGPs $\{F_n\}$ in $\mathcal{F}$ such that (i) $n^{1/2}\delta_{F_n} \to d \in \mathbb{R}^{r^*}$ or $\|n^{1/2}\delta_{F_n}\| \to \infty$; and (ii) G_{2,F_n}, Ω_{2,F_n} and M_{2,F_n} converges to $G_{2,F}$, $\Omega_{2,F}$ and $M_{2,F}$ for some $F \in \mathcal{F}$. [20]

For $k = 1, 2$ and any $F \in \mathcal{F}$, define

$$\Gamma_{k,F} = -\left(G'_{k,F}\Omega^{-1}_{k,F}G_{k,F}\right)^{-1}G'_{k,F}\Omega^{-1}_{k,F}. \tag{4.1}$$

Let $\mathcal{Z}_{2,F}$ denote a zero mean normal random vector with variance–covariance matrix $\Omega_{2,F}$ and $\mathcal{Z}_{1,F}$ denote its first r_1 components.

Lemma 4.1. *Suppose Assumptions 3.1 and 3.2 hold. Consider any sequence of DGPs $\{F_n\}$ such that $v_{F_n} \to v_F = (\mathrm{vec}(G_{2,F})', \mathrm{vech}(\Omega_{2,F})', \delta'_F)$ for some $F \in \mathcal{F}$, and $n^{1/2}\delta_{F_n} \to d$ for $d \in \mathbb{R}^{r^*}_{\infty}$.*

(a) *If $d \in \mathbb{R}^{r^*}$, then*

$$\begin{pmatrix} n^{1/2}(\widehat{\theta}_1 - \theta_{F_n}) \\ n^{1/2}(\widehat{\theta}_2 - \theta_{F_n}) \end{pmatrix} \to_D \begin{pmatrix} \xi_{1,F} \\ \xi_{2,F} \end{pmatrix} \equiv \begin{pmatrix} \Gamma_{1,F}\mathcal{Z}_{1,F} \\ \Gamma_{2,F}(\mathcal{Z}_{2,F} + d_0) \end{pmatrix},$$

where $d_0 = (\mathbf{0}_{1\times r_1}, d')'$.

[18] In the rest of the paper, we use $\{F_n\}$ to denote $\{F_n \in \mathcal{F} : n = 1, 2, \ldots\}$.

[19] Since $F_n \in \mathcal{F}$, by Assumption 3.2(ii), the sequence δ_{F_n} in (3.14) should satisfy $\|\delta_{F_n}\| \le C$ for any n.

[20] The requirement on the convergence of G_{2,F_n}, Ω_{2,F_n}, and M_{2,F_n} is not restrictive. Lemma B.7 in Appendix B.1 shows that the sequences G_{2,F_n}, Ω_{2,F_n}, and M_{2,F_n} have subsequences that converge to $G_{2,F}$, $\Omega_{2,F}$, and $M_{2,F}$, respectively, for some $F \in \mathcal{F}$. The general result on the lower and upper bounds of the asymptotic risk difference, Lemma B.14 in Appendix B.2, only requires to consider the subsequence $\{F_{p_n}\}$ such that $G_{2,F_{p_n}}$, $\Omega_{2,F_{p_n}}$, and $M_{2,F_{p_n}}$ are convergent, where $\{p_n\}$ is a subsequence of $\{n\}$. The asymptotic properties of the GMM estimators established in this section under the full sequence of DGPs $\{F_n\}$ holds trivially for its subsequence.

(b) *If* $\|d\| = \infty$, *then* $n^{1/2}(\widehat{\theta}_1 - \theta_{F_n}) \to_D \xi_{1,F}$ *and* $\|n^{1/2}(\widehat{\theta}_2 - \theta_{F_n})\| \to_p \infty$.

Given the joint asymptotic distribution of $\widehat{\theta}_1$ and $\widehat{\theta}_2$, it is straightforward to study $\widehat{\theta}(\omega) = (1-\omega)\widehat{\theta}_1 + \omega\widehat{\theta}_2$ if ω is deterministic. Following Lemma 4.1(a),

$$n^{1/2}\big(\widehat{\theta}(\omega) - \theta_{F_n}\big) \to_D \xi_F(\omega) \equiv (1-\omega)\xi_{1,F} + \omega\xi_{2,F} \tag{4.2}$$

for $n^{1/2}\delta_{F_n} \to d$, where $d \in \mathbb{R}^{r^*}$. In Section E of the Supplemental Appendix, a simple calculation shows that the asymptotic risk of $\widehat{\theta}(\omega)$ is minimized at the infeasible optimal weight

$$\omega_F^* \equiv \frac{\operatorname{tr}\big(\Upsilon(\Sigma_{1,F} - \Sigma_{2,F})\big)}{d_0'\big(\Gamma_{2,F} - \Gamma_{1,F}^*\big)'\Upsilon\big(\Gamma_{2,F} - \Gamma_{1,F}^*\big)d_0 + \operatorname{tr}\big(\Upsilon(\Sigma_{1,F} - \Sigma_{2,F})\big)}, \tag{4.3}$$

where Υ is the matrix specified in the loss function,

$$\Sigma_{k,F} \equiv \big(G_{k,F}'\Omega_{k,F}^{-1}G_{k,F}\big)^{-1} \quad \text{for } k = 1, 2 \text{ and } \Gamma_{1,F}^* \equiv [\Gamma_{1,F}, \mathbf{0}_{d_\theta \times r^*}]. \tag{4.4}$$

To gain some intuition, consider the case where $\Upsilon = I_{d_\theta}$ such that the MSE of $\widehat{\theta}(\omega)$ is minimized at ω_F^*. In this case, the infeasible optimal weight ω_F^* yields the ideal bias and variance trade off. However, the bias depends on d, which cannot be consistently estimated. Hence, ω_F^* cannot be consistently estimated. Our solution to this problem follows the popular approach in the literature which replaces d by an estimator whose asymptotic distribution is centered at d; see Liu (2015) and Charkhi, Claeskens, and Hansen (2016) for similar estimators in the least square estimation and maximum likelihood estimation problems, respectively.

The empirical analog of ω_F^* is constructed as follows. First, for $k = 1$ and 2, replace $\Sigma_{k,F}$ by its consistent estimator $\widehat{\Sigma}_k \equiv (\widehat{G}_k'\widehat{\Omega}_k^{-1}\widehat{G}_k)^{-1}$,[21] where

$$\begin{aligned} \widehat{G}_k &\equiv n^{-1}\sum_{i=1}^n g_{k,\theta}(W_i, \widehat{\theta}_1) \quad \text{and} \\ \widehat{\Omega}_k &\equiv n^{-1}\sum_{i=1}^n g_k(W_i, \widehat{\theta}_1)g_k(W_i, \widehat{\theta}_1)' - \overline{g}_k(\widehat{\theta}_1)\overline{g}_k(\widehat{\theta}_1)'. \end{aligned} \tag{4.5}$$

Note that $\widehat{G}_k$ and $\widehat{\Omega}_k$ are based on the conservative GMM estimator $\widehat{\theta}_1$. Hence they are consistent regardless of the degree of misspecification of the moment conditions in (1.3). Second, replace $(\Gamma_{2,F} - \Gamma_{1,F}^*)d_0$ by its asymptotically unbiased estimator $n^{1/2}(\widehat{\theta}_2 - \widehat{\theta}_1)$ because

$$n^{1/2}(\widehat{\theta}_2 - \widehat{\theta}_1) \to_D \big(\Gamma_{2,F} - \Gamma_{1,F}^*\big)(\mathcal{Z}_{2,F} + d_0), \tag{4.6}$$

for $d_0 = (\mathbf{0}_{1\times r_1}, d')'$ and $d \in \mathbb{R}^{r^*}$ following Lemma 4.1(a). Then the empirical optimal weight takes the form

$$\widetilde{\omega}_{\text{eo}} \equiv \frac{\operatorname{tr}\big(\Upsilon(\widehat{\Sigma}_1 - \widehat{\Sigma}_2)\big)}{n(\widehat{\theta}_2 - \widehat{\theta}_1)'\Upsilon(\widehat{\theta}_2 - \widehat{\theta}_1) + \operatorname{tr}\big(\Upsilon(\widehat{\Sigma}_1 - \widehat{\Sigma}_2)\big)}, \tag{4.7}$$

[21] The consistency of $\widehat{\Sigma}_k$ is proved in the proof of Lemma 4.2.

and the averaging GMM estimator takes the form

$$\widehat{\theta}_{\text{eo}} = (1-\widetilde{\omega}_{\text{eo}})\widehat{\theta}_1 + \widetilde{\omega}_{\text{eo}}\widehat{\theta}_2. \tag{4.8}$$

Next, we consider the asymptotic distribution of $\widehat{\theta}_{\text{eo}}$ under different degrees of misspecification.

Lemma 4.2. *Suppose that Assumptions 3.1–3.3 hold. Consider any sequence of DGPs $\{F_n\}$ such that $v_{F_n} \to v_F = (\text{vec}(G_{2,F})', \text{vech}(\Omega_{2,F})', \delta_F')$ for some $F \in \mathcal{F}$, and $n^{1/2}\delta_{F_n} \to d$ for $d \in \mathbb{R}^{r^*}_{\infty}$.*

(a) *If $d \in \mathbb{R}^{r^*}$, then*

$$\widetilde{\omega}_{\text{eo}} \to_D \overline{\omega}_F \equiv \frac{\text{tr}\big(\Upsilon(\Sigma_{1,F}-\Sigma_{2,F})\big)}{(\mathcal{Z}_{2,F}+d_0)'\big(\Gamma_{2,F}-\Gamma^*_{1,F}\big)'\Upsilon\big(\Gamma_{2,F}-\Gamma^*_{1,F}\big)(\mathcal{Z}_{2,F}+d_0)+\text{tr}\big(\Upsilon(\Sigma_{1,F}-\Sigma_{2,F})\big)}$$

and

$$n^{1/2}(\widehat{\theta}_{eo}-\theta_{F_n}) \to_D \overline{\xi}_F \equiv (1-\overline{\omega}_F)\xi_{1,F} + \overline{\omega}_F\xi_{2,F}.$$

(b) *If $\|d\| = \infty$, then $\widetilde{\omega}_{\text{eo}} \to_p 0$ and $n^{1/2}(\widehat{\theta}_{\text{eo}}-\theta_{F_n}) \to_D \xi_{1,F}$.*

To study the bounds of the asymptotic risk difference between $\widehat{\theta}_{\text{eo}}$ and $\widehat{\theta}_1$, it is important to take into account the data-dependent nature of $\widetilde{\omega}_{\text{eo}}$. Unlike $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$, the randomness in $\widetilde{\omega}_{\text{eo}}$ is non-negligible in the mild misspecification case (a) of Lemma 4.2. In consequence, $\widehat{\theta}_{\text{eo}}$ does not achieve the same bounds of asymptotic risk difference as the ideal averaging estimator $(1-\omega^*_F)\widehat{\theta}_1 + \omega^*_F\widehat{\theta}_2$ does. Nevertheless, below we show that $\widehat{\theta}_{\text{eo}}$ is insured against potentially misspecified moments because it uniformly dominates $\widehat{\theta}_1$.

5. Bounds of asymptotic risk difference under misspecification

In this section, we study the bounds of the asymptotic risk difference defined in (3.12). Note that the asymptotic distributions of $\widehat{\theta}_1$ and $\widehat{\theta}_{\text{eo}}$ in Lemmas 4.1 and 4.2 only depend on d, $G_{2,F}$, and $\Omega_{2,F}$. For notational convenience, define

$$h_{F,d} = \big(d', \text{vec}(G_{2,F})', \text{vech}(\Omega_{2,F})'\big) \tag{5.1}$$

for any $F \in \mathcal{F}$ and any $d \in \mathbb{R}^{r^*}_{\infty}$. For the mild misspecification case, define the parameter space of $h_{F,d}$ as

$$H = \big\{h_{F,d} : d \in \mathbb{R}^{r^*} \text{ and } F \in \mathcal{F} \text{ with } \delta_F = 0_{r^*\times 1}\big\}, \tag{5.2}$$

where δ_F is defined by (1.5) for a given F.

Theorem 5.1. *Suppose that Assumptions 3.1–3.3 hold. The bounds of the asymptotic risk difference satisfy*

$$\text{Asy}\overline{RD}(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1) = \max\Big\{\sup_{h\in H}\big[g(h)\big], 0\Big\},$$

$$\text{Asy}\underline{RD}(\widehat{\theta}_{\text{eo}}, \widehat{\theta}_1) = \min\Big\{\inf_{h\in H}\big[g(h)\big], 0\Big\},$$

where $g(h) \equiv \mathbb{E}[\overline{\xi}_F' \Upsilon \overline{\xi}_F - \xi_{1,F}' \Upsilon \xi_{1,F}]$, $\xi_{1,F}$ *and* $\overline{\xi}_F$ *are given in Lemma 4.1 and Lemma 4.2, respectively, and the expectation is taken under the joint normal distribution with mean zero and variance–covariance matrix* $\Omega_{2,F}$.

The upper (or lower) bound of the asymptotic risk difference is determined by the maximum between $\sup_{h\in H}[g(h)]$ and zero (or the minimum between $\inf_{h\in H}[g(h)]$ and zero), where $\sup_{h\in H}[g(h)]$ (or $\inf_{h\in H}[g(h)]$) is related to the mildly misspecified DGPs and the zero component is associated with the severely misspecified DGPs. Since the GMM averaging estimator has the same asymptotic distribution as the conservative GMM estimator $\widehat{\theta}_1$ under the severely misspecified DGPs, their asymptotic risk difference is zero.

To show that $\widehat{\theta}_{\mathrm{eo}}$ uniformly dominates $\widehat{\theta}_1$ following (3.13), Theorem 5.1 implies that it is sufficient to show that $\inf_{h\in H}[g(h)] < 0$ and $\sup_{h\in H}[g(h)] \leq 0$. We can investigate $\inf_{h\in H} g(h)$ and $\sup_{h\in H} g(h)$ by simulating $g(h)$. In practice, we replace $G_{2,F}$ and $\Omega_{2,F}$ by their consistent estimators and plot $g(h)$ as a function of d. Even if the uniform dominance condition does not hold, $\min\{\inf_{h\in H}[g(h)], 0\}$ and $\max\{\sup_{h\in H}[g(h)], 0\}$ quantify the most- and least-favorable scenarios for the averaging estimator.

Theorem 5.2. *Let* $A_F \equiv \Upsilon(\Sigma_{1,F} - \Sigma_{2,F})$ *for any* $F \in \mathcal{F}$. *Suppose that Assumptions 3.1–3.3 hold. If* $\mathrm{tr}(A_F) > 0$ *and* $\mathrm{tr}(A_F) \geq 4\rho_{\max}(A_F)$ *for any* $F \in \mathcal{F}$ *with* $\delta_F = 0$, *we have*

$$\mathrm{Asy}\underline{RD}(\widehat{\theta}_{\mathrm{eo}}, \widehat{\theta}_1) < 0 \quad \textit{and} \quad \mathrm{Asy}\overline{RD}(\widehat{\theta}_{\mathrm{eo}}, \widehat{\theta}_1) = 0.$$

Thus, $\widehat{\theta}_{\mathrm{eo}}$ *uniformly dominates* $\widehat{\theta}_1$.

Theorem 5.2 indicates that: (i) there exists $\varepsilon_1 < 0$ and some finite integer n_{ε_1} such that the minimum risk difference between $\widehat{\theta}_{\mathrm{eo}}$ and $\widehat{\theta}_1$ is less than ε_1 for any n larger than n_{ε_1}; (ii) for any $\varepsilon_2 > 0$, there exists a finite integer n_{ε_2} such that the maximum risk difference between $\widehat{\theta}_{\mathrm{eo}}$ and $\widehat{\theta}_1$ is less than ε_2 for any n larger than n_{ε_2}. Pre-test estimators fail to satisfy both properties (i) and (ii) above at the same time. Take the pre-test estimator based on the J-test, for example,[22] and consider three scenarios: (a) the critical value is fixed for any sample size; (b) the critical value diverges to infinity; and (c) the critical value converges to zero. In the pointwise asymptotic framework, the J-test based on the critical values in (a), (b), and (c) leads to inconsistent (but conservative) model selection, consistent model selection, and no model selection results, respectively. The pre-test estimator based on the J-test violates property (ii) in scenarios (a) and (b), and violates property (i) in scenario (c).

Different from the finite-sample results for the JS estimator established for the Gaussian location model, our comparison of the two estimators $\widehat{\theta}_{\mathrm{eo}}$ and $\widehat{\theta}_1$ is based on the asymptotic bounds of the risk difference. For a given sample size n, we do not provide results on this asymptotic approximation error and, therefore, our results do not state how the finite-sample upper bound $\overline{RD}_n(\widehat{\theta}_{\mathrm{eo}}, \widehat{\theta}_1; \zeta)$ approaches to zero as $n \to \infty$ and then $\zeta \to \infty$ (e.g., from above or from below). For the Gaussian location model,

[22] See Section F in the Supplemental Appendix for definition and analysis of this estimator.

the asymptotically uniform dominance here is weaker than the classical finite-sample results established for the JS estimator. However, the asymptotic results here apply to general nonlinear econometric models with nonnormal random variables.[23]

To shed light on the sufficient conditions in Theorem 5.2, let us consider a scenario similar to the JS estimator: $\Sigma_{1,F} = \sigma_{1,F}^2 I_{d_\theta}$, $\Sigma_{2,F} = \sigma_{2,F}^2 I_{d_\theta}$, and $\Upsilon = I_{d_\theta}$. In this case, the sufficient conditions become $\sigma_{1,F} > \sigma_{2,F}$ and $d_\theta \geq 4$. The first condition $\mathrm{tr}(A_F) > 0$, which is reduced to $\sigma_{1,F} > \sigma_{2,F}$, requires that the additional moments $\mathbb{E}_F[g^*(W_i, \theta_F)] = 0$ are nonredundant in the sense that they lead to a more efficient estimator of θ_F. The second condition $\mathrm{tr}(A_F) \geq 4\rho_{\max}(A_F)$, which is reduced to $d_\theta \geq 4$, requires that we are interested in the total risk of several parameters rather than that of a single one. In a more general case where $\Sigma_{1,F}$ and $\Sigma_{2,F}$ are not proportional to the identity matrix, the sufficient conditions are reduced to $\Sigma_{1,F} > \Sigma_{2,F}$ and $d_\theta \geq 4$ under the choice $\Upsilon = (\Sigma_{1,F} - \Sigma_{2,F})^{-1}$, which rescales $\widehat{\theta}$ by the variance reduction $\Sigma_{1,F} - \Sigma_{2,F}$. In a simple linear IV model (Example 3.1) where Z_i^* is independent of $Z_{1,i}$ and the regression error U_i is homoskedastic conditional on the IVs, $\Sigma_{1,F} > \Sigma_{2,F}$ requires that $\mathbb{E}_{F^*}[Z_i^* X_i']$ and $\mathbb{E}_{F^*}[Z_i^* Z_i^{*\prime}]$ both have full rank.

6. Simulation studies

In this section, we investigate the finite sample performance of our averaging GMM estimator in linear IV models. In addition to the empirical optimal weight $\widetilde{\omega}_{\mathrm{eo}}$, we consider another averaging estimator based on the JS type of weight. Define the positive part of the JS weight:[24]

$$\omega_{\mathrm{JS}} = 1 - \left(1 - \frac{\mathrm{tr}(\widehat{A}) - 2\rho_{\max}(\widehat{A})}{n(\widehat{\theta}_2 - \widehat{\theta}_1)' \Upsilon (\widehat{\theta}_2 - \widehat{\theta}_1)}\right)_+, \tag{6.1}$$

where $(x)_+ = \max\{0, x\}$ and $\widehat{A}$ is the estimator of A_F using $\widehat{\Sigma}_k$ for $k = 1, 2$. In the simulation study of this paper, we consider an alternative averaging estimator with the restricted JS weight

$$\omega_{R,\mathrm{JS}} = (\omega_{\mathrm{JS}})_+. \tag{6.2}$$

By construction, $\omega_{\mathrm{JS}} \leq 1$ and $0 \leq \omega_{R,\mathrm{JS}} \leq 1$. We compare the finite-sample (truncated and untruncated) MSEs of our proposed averaging estimator with the empirical optimal weight, the JS type of averaging estimator with the restricted weight in (6.2), the conservative GMM estimator $\widehat{\theta}_1$, and the pre-test GMM estimator based on the J-test. The finite-sample MSE of the conservative GMM estimator is normalized to 1. That is, we report the ratios of various MSEs to the MSE of the conservative GMM estimator and call these ratios as relative MSEs. Three different simulation designs are considered in this section.

[23] In Section A of the Appendix, we show that the averaging GMM estimator has similar finite sample dominance results in the Gaussian location model.

[24] This formula is a GMM analog of the generalized JS-type shrinkage estimator in Hansen (2016) for parametric models. The shrinkage scalar τ is set to $\mathrm{tr}(\widehat{A}) - 2\rho_{\max}(\mathrm{tr}(\widehat{A}))$ in a fashion similar to the original JS estimator.

6.1 *Simulation model 1*

We consider a linear regression model with i.i.d. observed data

$$W_i = \left(Y_i, X_{1,i}, \dots, X_{6,i}, Z_{1,i}, \dots, Z_{12,i}, Z^*_{1,i}, \dots, Z^*_{6,i}\right)' \quad \text{for } i = 1, \dots, n, \tag{6.3}$$

where Y is the dependent variable, $(X_1, \dots, X_6)$ are 6 endogenous regressors, $(Z_1, \dots, Z_{12})$ are 12 valid IVs, and $(Z^*_1, \dots, Z^*_6)$ are 6 potentially invalid IVs. The data are generated as follows. The regression model is

$$Y = \sum_{j=1}^{6} \theta_j X_j + u, \tag{6.4}$$

where $(\theta_1, \dots, \theta_6)$ is set to $2.5 \times \mathbf{1}_{1\times 6}$ and X_j is generated by

$$X_j = 2^{-1}(Z_j + Z_{j+6}) + Z_{j+12} + \varepsilon_j \quad \text{for } j = 1, \dots, 6. \tag{6.5}$$

We first draw $(Z_1, \dots, Z_{18}, \varepsilon_1, \dots, \varepsilon_6, u^*)'$ from normal distribution with mean zero and variance–covariance matrix $\operatorname{diag}(I_{18\times 18}, \Sigma_{7\times 7})$ where

$$\Sigma_{7\times 7} = \begin{pmatrix} I_{6\times 6} & 0.25 \times \mathbf{1}_{6\times 1} \\ 0.25 \times \mathbf{1}_{1\times 6} & 1 \end{pmatrix}. \tag{6.6}$$

We consider two designs for generating the structural error u in (6.4). The first design (S1 hereafter) has non-Gaussian errors. Draw η from exponential distribution with mean 1 and η is independent of $(Z_1, \dots, Z_{18}, \varepsilon_1, \dots, \varepsilon_6, u^*)$. Generate the structural error

$$u = \left(u^* + \eta_0\right)/2, \tag{6.7}$$

where η_0 is the demeaned version of η to ensure that the mean of u is zero. The second design (S2 hereafter) has normal error

$$u = u^*. \tag{6.8}$$

The potentially invalid IVs are generated by

$$Z^*_j = \left(1 - c_j^2\right)^{1/2} Z_{j+12} + c_j(\varepsilon_j + u), \tag{6.9}$$

where $c_j \in [0, 1]$ for $j = 1, \dots, 6$. In this simulation study, we consider different DGPs by choosing various values for $c = (c_1, \dots, c_6)$ where $c_j \in [0, 1]$ for $j = 1, \dots, 6$. Therefore,

$$\mathbb{E}\left[u Z^*_j\right] = \begin{cases} 5c_j/8 & \text{under (6.7)}, \\ 5c_j/4 & \text{under (6.8)}. \end{cases} \tag{6.10}$$

From the above expression, we see that Z^*_j is a valid IV if c_j is zero while increasing c_j to 1 will enlarge the correlation coefficient between Z^*_j and u, and hence the endogeneity of Z^*_j.

Given the sample size n, different DGPs of the simulated data $\{W_i : i = 1, \ldots, n\}$ are employed in the simulation study by changing the values of $(c_1, \ldots, c_6)$. We consider

$$c_j = c_0 \omega_j \quad \text{for } j = 1, \ldots, 6 \tag{6.11}$$

where c_0 is a scalar that takes values on the grid points between 0 and 1 with the grid length 0.02, and $(\omega_1, \ldots, \omega_6)$ is parametrized in two different ways. In the first one, we set $\omega_j = 0$ or 1 for $j = 1, \ldots, 6$ and rule out the case that $\omega_j = 0$ for all j (since this is the same as the case which sets $c_0 = 0$). In the second one, we consider the polar transformation and set

$$\begin{aligned} \omega_1 &= \sin(\alpha_1)\sin(\alpha_2)\sin(\alpha_3)\sin(\alpha_4)\sin(\alpha_5), \\ \omega_j &= \cos(\alpha_{j-1})\sin(\alpha_j) \times \cdots \times \sin(\alpha_5) \quad \text{for } j = 2, \ldots, 5, \\ \omega_6 &= \cos(\alpha_5), \end{aligned} \tag{6.12}$$

where $\alpha_1 \in \{\pi/4, 3\pi/4, 5\pi/4, 7\pi/4\}$ and $\alpha_j \in \{\pi/4, 3\pi/4\}$ for $j = 2, \ldots, 5$. Therefore, there are 127 different values for $(\omega_1, \ldots, \omega_6)$ for each of the 51 different values of c_0. For each DGP, we consider sample size $n = 50, 100, 250, 500, 1000$ and use 10,000 simulation repetitions.

Given the sample size and the value of c_0, we report the minimum and the maximum of the 127 values of the finite sample relative MSEs for each estimator, and the weight $\widetilde{\omega}_{\text{eo}}$ in our averaging estimator in the DGP with the maximum relative MSE. Given each sample size, the maximum/minimum finite sample relative MSE and the weight are plotted as functions of c_0; see Figure 2 for S1 and Figure 3 for S2. In each figure, the left three panels and the right three panels include the results with sample size $n = 100$ and 500, respectively.[25] For each sample size, we also report the upper bound and the lower bound of the finite sample relative MSEs (among all 127 × 51 DGPs) of the averaging estimators and the pre-test estimator in Table 1.[26]

Our findings in the simulation designs S1 and S2 are summarized as follows. First, in both Figure 2 and Figure 3, we see that the minimum relative MSE of the averaging GMM estimator $\widehat{\theta}_{\text{eo}}$ is smaller than 1 (which is the normalized finite sample MSE of the conservative GMM estimator $\widehat{\theta}_1$) for all c_0 considered in both simulation designs. The maximum relative MSE of $\widehat{\theta}_{\text{eo}}$ is smaller than 1 when c_0 is small and approaches 1 when c_0 is close to 1. Table 1 provides detailed information on the lower and upper bounds of the relative MSE of $\widehat{\theta}_{\text{eo}}$. In both simulation designs S1 and S2, the lower bound stays far below 1 while the upper bound approaches 1 with increasing sample size. These results are predicted by our theory because the key sufficient condition is satisfied in both

[25] We only report the untruncated MSEs with $n = 100$ and $n = 500$ here. The untruncated MSEs in S1 and S2 with $n = 50$, 250, and 1000 can be found in Figure C.1 and Figure C.2 in Section C of the Appendix, and the truncated MSEs ($\zeta = 1000$) in S1 and S2 with $n = 50$, 100, 250, 500, and 1000 can be found in Figure G.1, Figure G.2, Figure G.4, and Figure G.5 in Section G of the Supplemental Appendix. The simulation results on truncated MSEs are very similar to what we get without truncation. The maximum finite sample bias and finite sample variance for each c_0 are reported in Figure C.4 and Figure C.5 in Section C of the Appendix.

[26] The upper bound and the lower bound of the finite sample relative truncated MSEs are reported in Table G.1 in Section G of the Supplemental Appendix.

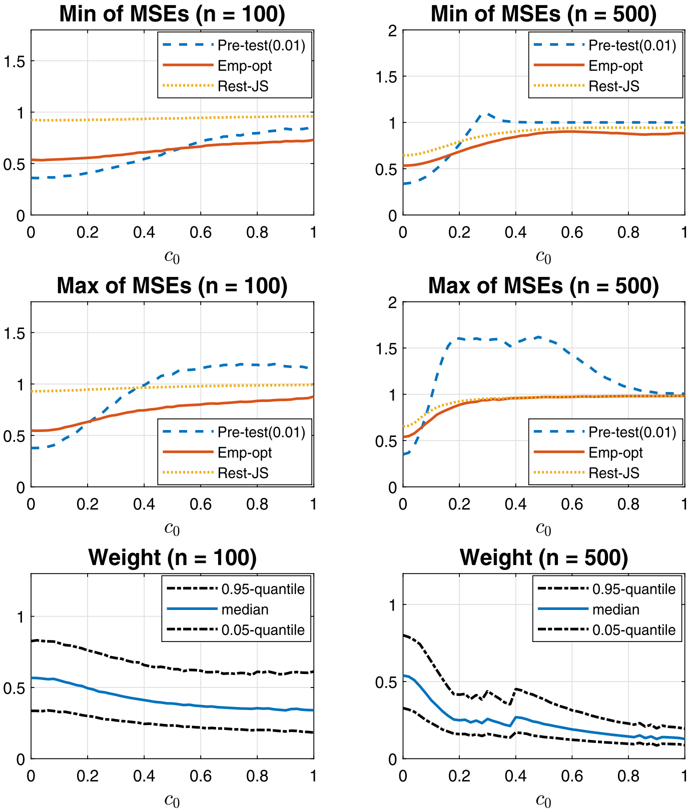

FIGURE 2. Finite sample MSEs of the pre-test and averaging GMM estimators in S1. *Note*: "Pre-test(0.01)" refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; "Emp-opt" refers to the averaging GMM estimator based on the empirical optimal weight; "Rest-JS" refers to the averaging estimators based on the restricted James–Stein weight, respectively.

S1 and S2.[27] Second, the pre-test GMM estimator has nonshrinking maximum relative MSE in S1 or S2 and, therefore, it fails to dominate the conservative GMM estimator $\widehat{\theta}_1$ in the asymptotic sense. For example, when $n = 500$, the pre-test GMM estimator in S1 has relative MSE above 1.5 when c_0 is between 0.2 and 0.4. From Table 1, we see that the upper bound of its relative MSE does not converge to 1 with increasing sample size. It stays around 1.61 and 1.69 in S1 and S2, respectively, when the sample size is large (e.g., $n = 500$ or 1000). Third, comparing the two averaging estimators, we find that the restricted JS estimator does not reduce the MSE as much as the averaging estimator based on $\widetilde{\omega}_{\text{eo}}$. Fourth, the weight $\widetilde{\omega}_{\text{eo}}$ becomes close to zero when c_0 is close to 1 for large n, which is clearly illustrated by the simulation with $n = 500$ in both S1 and S2. Last, the maximum relative MSE of the pre-test GMM estimator may show multiple peaks in Figure 2 and Figure 3, because given c_0 the Euclidean norm of $(c_1, \ldots, c_6)$ may be different

[27] It is easy to show that when $\delta_F = 0$ and Υ is the identity matrix, we have $\text{tr}(A_F) = 4$ and $\text{tr}(A_F) - 4\rho_{\max}(A_F) = 4/3$ for S1 and $\text{tr}(A_F) = 8$ and $\text{tr}(A_F) - 4\rho_{\max}(A_F) = 8/3$ for S2.

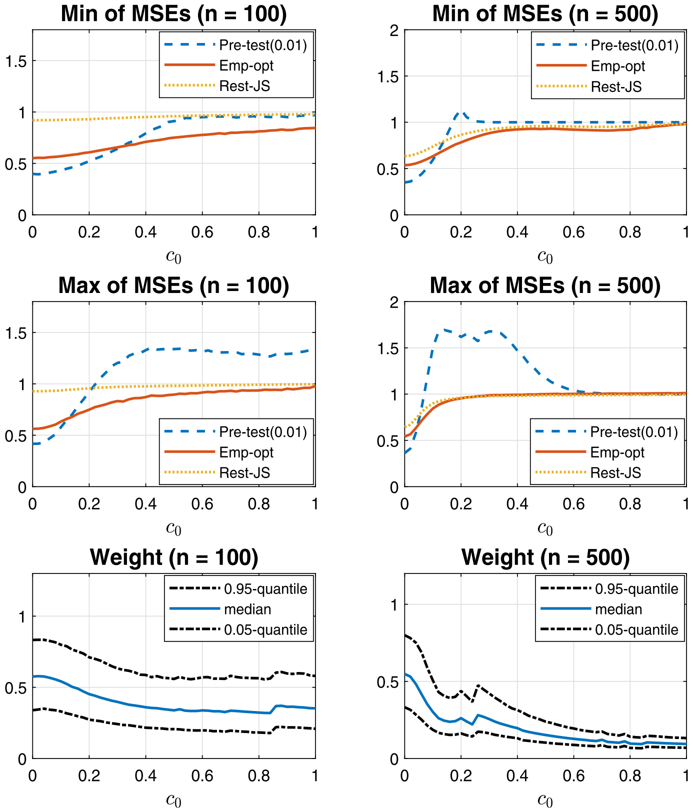

FIGURE 3. Finite sample MSEs of the pre-test and averaging GMM estimators in S2. *Note*: “Pre-test(0.01)” refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; “Emp-opt” refers to the averaging GMM estimator based on the empirical optimal weight; “Rest-JS” refers to the averaging estimators based on the restricted James–Stein weight, respectively.

under the two different parametrizations of $(\omega_1, \dots, \omega_6)$. In the polar transformation, the Euclidean norm of $(c_1, \dots, c_6)$ is c_0. However, the Euclidean norm of $(c_1, \dots, c_6)$ is $c_0(\omega_1 + \dots + \omega_6)^{1/2}$ in the other design and it may take 5 different values when we set $\omega_j = 0$ or 1 for $j = 1, \dots, 6$ and rule out the case that $\omega_j = 0$ for all j. This also explains the kinks of the weight $\widetilde{\omega}_{\text{eo}}$ in the averaging estimator associated with the maximum MSE.

6.2 *Simulation model* 2

In this subsection, we investigate the finite sample properties of the pre-test GMM estimator and the averaging GMM estimators when the key uniform dominance condition in Theorem 5.2 does not hold. In this simulation design, the structural equation takes the same form as (6.4) with $(\theta_1, \dots, \theta_6) = 2.5 \times \mathbf{1}_{1\times 6}$, but the regressors X_j $(j = 1, \dots, 6)$ are generated in a different way. We draw i.i.d. random vectors $(Z_1, \dots, Z_{13}, \varepsilon_1, \dots, \varepsilon_5, u)$ from normal distribution with mean zero and variance–

Table 1. The lower and upper bounds of the finite sample relative MSEs.

		Design S1		Design S2		Design S3	
		Lower	Upper	Lower	Upper	Lower	Upper
$n = 50$	$\hat{\theta}_{\text{oe}}$	0.5732	0.7968	0.6113	0.8980	0.9302	1.0012
	$\hat{\theta}_{\text{JS}}$	0.9755	0.9959	0.9776	0.9978	0.9995	1.0003
	$\hat{\theta}_{\text{pret}}$	0.4424	0.9574	0.5057	1.0973	1.0324	1.4283
$n = 100$	$\hat{\theta}_{\text{oe}}$	0.5325	0.8789	0.5513	0.9781	0.9733	1.0040
	$\hat{\theta}_{\text{JS}}$	0.9208	0.9911	0.9202	0.9956	0.9996	1.0002
	$\hat{\theta}_{\text{pret}}$	0.3586	1.1940	0.3937	1.3539	0.9990	1.4709
$n = 250$	$\hat{\theta}_{\text{oe}}$	0.5316	0.9587	0.5384	1.0118	0.9720	1.0079
	$\hat{\theta}_{\text{JS}}$	0.7591	0.9787	0.7506	0.9923	0.9999	1.0000
	$\hat{\theta}_{\text{pret}}$	0.3360	1.5106	0.3598	1.6392	0.9753	1.4394
$n = 500$	$\hat{\theta}_{\text{oe}}$	0.5331	0.9846	0.5355	1.0112	0.9700	1.0096
	$\hat{\theta}_{\text{JS}}$	0.6443	0.9823	0.6359	0.9953	1.0000	1.0000
	$\hat{\theta}_{\text{pret}}$	0.3368	1.6196	0.3495	1.6937	0.9562	1.4236
$n = 1000$	$\hat{\theta}_{\text{oe}}$	0.5335	0.9934	0.5341	1.0082	0.9681	1.0119
	$\hat{\theta}_{\text{JS}}$	0.5803	0.9890	0.5737	0.9978	1.0000	1.0000
	$\hat{\theta}_{\text{pret}}$	0.3395	1.6433	0.3451	1.6864	0.9473	1.3953

Note: 1. $\hat{\theta}_{\text{JS}}$ and $\hat{\theta}_{\text{pret}}$ denote the GMM averaging estimator based on the weight in (6.1) and the pre-testing GMM estimator based on J-test with nominal size 0.01, respectively; 2. the "Upper" and "Lower" refer to the upper bound and the lower bound of the finite sample relative MSEs among all DGPs considered in the simulation design given the sample size.

covariance matrix $\text{diag}(I_{13\times 13}, \Sigma_{6\times 6})$, where

$$\Sigma_{6\times 6} = \begin{pmatrix} I_{5\times 5} & 0.25 \times \mathbf{1}_{5\times 1} \\ 0.25 \times \mathbf{1}_{1\times 5} & 1 \end{pmatrix}. \tag{6.13}$$

The observed data are $W = (Y, X_1, \dots, X_6, Z_6, Z_7, Z_8, Z_1^*, \dots, Z_5^*)$, where $(X_1, \dots, X_5)$ are exogenous regressors and X_6 is an endogenous regressor, $(X_1, \dots, X_5, Z_6, Z_7, Z_8)$ are valid IVs and $(Z_1^*, \dots, Z_5^*)$ are potentially invalid IVs. The exogenous variables are generated by

$$\begin{aligned} X_j &= 3^{-\frac{1}{2}}(Z_j + Z_{j+1} + Z_{j+8}), \quad \text{for } j = 1, \dots, 4, \\ X_5 &= 3^{-\frac{1}{2}}(Z_5 + Z_1 + Z_{13}). \end{aligned} \tag{6.14}$$

The endogenous variable X_6 is generated by

$$X_6 = 2^{-1}\sum_{j=6}^{8} Z_j + 10^{-1/2}\sum_{j=1}^{5}(Z_{j+8} + \varepsilon_j). \tag{6.15}$$

The potentially invalid IVs are generated by

$$Z_j^* = \left(1 - c_j^2\right)^{1/2} Z_{j+8} + c_j(\varepsilon_j + u) \quad \text{for } j = 1, \dots, 5. \tag{6.16}$$

Note that the key sufficient condition in Theorem 5.2 is not satisfied for this design.[28] We call the simulation design in this subsection as S3.

Given the sample size n, we consider different DGPs of the simulated data $\{W_i : i = 1, \dots, n\}$ by changing the values of $(c_1, \dots, c_5)$. We consider the following parametrization:

$$c_j = c_0 \omega_j \quad \text{for } j = 1, \dots, 5, \tag{6.17}$$

where c_0 is a scalar that takes values on the grid points between 0 and 1 with the grid length 0.02, $(\omega_1, \dots, \omega_5)$ is parametrized in two different ways. In the first one, we set $\omega_j = 0$ or 1 for $j = 1, \dots, 5$ and rule out the case that $\omega_j = 0$ for all j (since this is the same as the case which sets $c_0 = 0$). In the second one, we consider the polar transformation and set

$$\begin{aligned} \omega_1 &= \sin(\alpha_1)\sin(\alpha_2)\sin(\alpha_3)\sin(\alpha_4), \\ \omega_j &= \cos(\alpha_{j-1})\sin(\alpha_j) \times \cdots \times \sin(\alpha_4) \quad \text{for } j = 2, \dots, 4, \\ \omega_5 &= \cos(\alpha_4), \end{aligned} \tag{6.18}$$

where $\alpha_1 \in \{\pi/4, 3\pi/4, 5\pi/4, 7\pi/4\}$ and $\alpha_j \in \{\pi/4, 3\pi/4\}$ for $j = 2, \dots, 5$. Therefore, there are 63 different values for $(\omega_1, \dots, \omega_5)$ for each of the 51 different values of c_0. For each DGP, we consider sample size $n = 50$, 100, 250, 500, 1000, and use 10,000 simulation repetitions.

Given the sample size and the value of c_0, we report the minimum and maximum of the 63 values of the finite sample relative MSEs for each estimator, and the weight $\widetilde{\omega}_{\mathrm{eo}}$ in our averaging estimator in the DGP with maximum relative MSE. Given each sample size, the maximum/minimum finite sample relative MSE and the weight are plotted as functions of c_0; see Figure 4.[29] For each sample size, the upper bound and the lower bound of the finite sample relative MSEs (among all 127×51 DGPs) of the averaging estimators and the pre-test estimator in this simulation design are also reported in Table 1.

Our findings in this simulation design are summarized as follows. First, compared to the conservative GMM estimator, the improvement of the pre-test GMM estimator or the averaging GMM estimator is small even when all the IVs Z_j^* $(j = 1, \dots, 5)$ are valid. This is because there is only one endogenous regressor and the improvement of using Z_j^* $(j = 1, \dots, 5)$ is mainly through the estimation of its coefficient. Second, both the pre-test GMM estimator and our averaging GMM estimator fail to dominate the conservative GMM estimator. However, the overall performance of the averaging GMM estimator is better than the pre-test GMM estimator. For example, when the sample size is 500, the maximum MSE of the pre-test GMM estimator is 1.4 times of that of the conservative

[28] It is easy to show that, when $\delta_F = 0$, we have $\mathrm{tr}(A_F) = 0.4916$ and $\mathrm{tr}(A_F) - 4\rho_{\max}(A_F) = -1.4748 < 0$.

[29] We only report the untruncated MSEs with $n = 100$ and $n = 500$ here. The untruncated MSEs in S3 with $n = 50$, 250 and 1000 can be found in Figure C.3 in Section C of the Appendix, and the truncated MSEs $(\zeta = 1000)$ in S3 with $n = 50$, 100, 250, 500, and 1000 can be found in Figure G.3 and Figure G.6 in Section G of the Supplemental Appendix. The simulation results on truncated MSEs are very similar to what we get without truncation. The maximum finite sample bias and finite sample variance for each c_0 are reported in Figure C.6 in Section C of the Appendix.

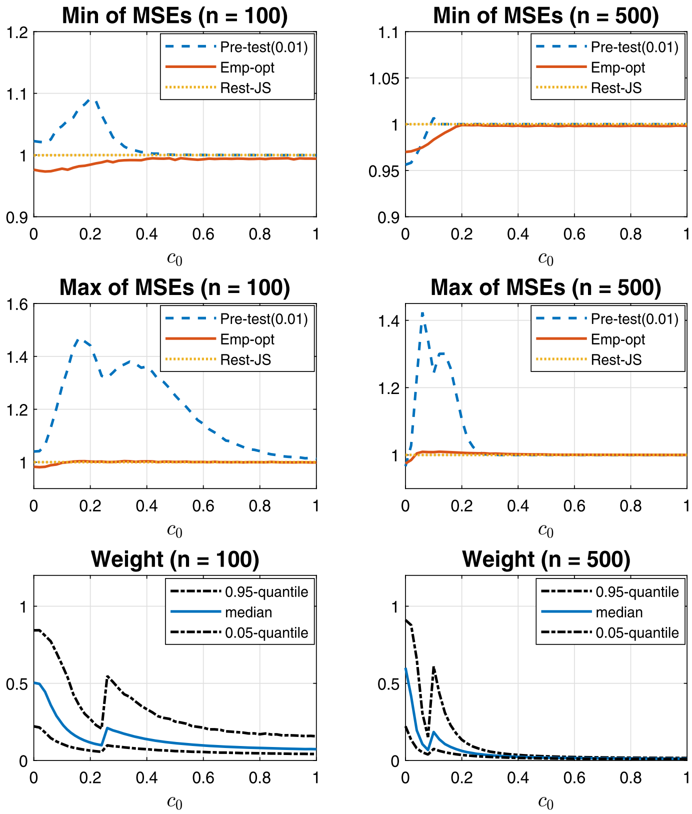

FIGURE 4. Finite sample MSEs of the pre-test and averaging GMM estimators in S3. *Note*: "Pre-test(0.01)" refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; "Emp-opt" refers to the averaging GMM estimator based on the empirical optimal weight; "Rest-JS" refers to the averaging estimators based on the restricted James–Stein weight, respectively.

GMM estimator. In contrast, the maximum MSE of our averaging GMM estimator is only slightly higher than (1.01 times of) that of the conservative GMM estimator. Third, the MSE of the JS-type averaging estimator is identical to the conservative GMM estimator even when all the IVs Z_j^* $(j = 1, \ldots, 5)$ are valid. Therefore, this estimator performs the same as the conservative GMM estimator. Fourth, as c_0 goes to 1, the weight $\widetilde{\omega}_{\mathrm{eo}}$ goes to zero for large sample size, which is well illustrated by the simulation with $n = 500$. Last, the maximum MSE of the pre-test GMM estimator and the weight $\widetilde{\omega}_{\mathrm{eo}}$ in our averaging estimator may show multiple peaks for the same reason explained in the previous subsection.

7. CONCLUSION

This paper studies the averaging GMM estimator that combines the conservative estimator and the aggressive estimator with a data-dependent weight. The averaging weight is the sample analog of an optimal nonrandom weight. We provide a sufficient class of drifting DGPs under which the pointwise asymptotic results combine to yield uniform approximations to the finite-sample risk difference between two estimators. Using this asymptotic approximation, we show that the proposed averaging GMM estimator uniformly dominates the conservative GMM estimator for quadratic loss functions such as the mean square errors.

Inference based on the averaging estimator is an interesting and challenging problem. As pointed out in Pötscher (2006), the finite sample density of the averaging estimator cannot be consistently estimated, which implies that directly applying an estimator of the finite-sample density may not yield uniformly valid inference. In addition to the uniform validity, a desirable confidence set should have smaller volume than that obtained from the conservative moments alone. We leave the inference issue to future investigation.

Appendix A: Illustration in Gaussian location model

This section shows that in a Gaussian location model, the averaging GMM estimator dominates the conservative GMM estimator in finite samples, that is, it exhibits the JS phenomenon.

Suppose that we have one observation $(X', Y')'$ from the normal distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ \theta + d \end{pmatrix}, \sigma^2 I_{2k} \right), \tag{A.1}$$

where σ^2 is a known positive value, θ and d are $k \times 1$ vectors, and I_{2k} is a $2k \times 2k$ identity matrix ($k \geq 3$). It is clear that $d_\theta = k$ in model (A.1). We are interested in estimating θ.

Green and Strawderman (1991) considered the same model defined in (A.1). They propose the following JS type of estimator:

$$\widehat{\theta}_{\mathrm{GS}} = X - \frac{\tau\sigma^2}{(X - Y)'(X - Y)}(X - Y), \tag{A.2}$$

where τ is a real constant in $(0, 2(k-2))$. Apparently, the above estimator $\widehat{\theta}_{\mathrm{GS}}$ is an averaging estimator which combines an unbiased estimator X with a biased estimator Y with weight $\tau\sigma^2\|X - Y\|^{-2}$ on the biased estimator. Green and Strawderman (1991) showed that when $k \geq 3$, $\widehat{\theta}_{\mathrm{GS}}$ has smaller MSE than the unbiased estimator X (which is the MLE of θ) for any $\theta \in \mathbb{R}^k$, any $d \in \mathbb{R}^k$ and any $\sigma^2 > 0$, and hence it uniformly dominates the MLE of θ.

Kim and White (2001), Judge and Mittelhammer (2004), and Mittelhammer and Judge (2005) proposed averaging estimators which shrink the (asymptotic) unbiased estimator toward the biased estimator in semiparametric regression models. These papers show the dominance of the averaging estimator over the (asymptotic) unbiased estimator in the Gaussian location models using the joint normal distribution of the unbiased and biased estimators. In these papers, X and Y are the unbiased and biased estimators, respectively, with general variance–covariance matrix that allows for correlation between X and Y. The averaging estimator proposed in Kim and White (2001) is

$$\widehat{\theta}_{KW} = X - \left(c_1 + \frac{c_2}{(X-Y)'(X-Y)}\right)(X-Y), \tag{A.3}$$

where c_1 and c_2 constants. When $k \geq 5$, Kim and White (2001) showed that there exist optimal values for c_1 and c_2 such that $\widehat{\theta}_{KW}$ dominates the unbiased estimator X. In the semiparametric setting, they show that these optimal values can be consistently estimated when $\mathbb{E}[Y] = \theta$. In Judge and Mittelhammer (2004) and Mittelhammer and Judge (2005), the averaging estimator takes the same form as $\widehat{\theta}_{\mathrm{GS}}$ in (A.2) except $\tau\sigma^2$ is replaced by a constant. They show that when $k \geq 5$, there exists an optimal constant under which their averaging estimator dominates the unbiased estimator X. They provide an approximator of the infeasible constant and show that the approximator can be consistently estimated.

We next consider our averaging GMM estimator. Let Y be the $k \times k$ identity matrix. The conservative GMM estimator $\widehat{\theta}_1 = X$ has risk $\sigma^2 \operatorname{tr}(YI_k) = \sigma^2 k$. On the other hand, the aggressive GMM estimator is $\widehat{\theta}_2 = (X+Y)/2$, which has risk $\sigma^2 k/2 + \|d\|^2/4$. The empirical optimal weight defined in (4.7) becomes

$$\widetilde{\omega}_{\mathrm{eo}} = \frac{2k\sigma^2}{2k\sigma^2 + (X-Y)'(X-Y)}, \tag{A.4}$$

which together with the conservative and aggressive GMM estimators leads to the averaging GMM estimator

$$\widehat{\theta}_{\mathrm{eo}} = X - \frac{k\sigma^2}{2k\sigma^2 + (X-Y)'(X-Y)}(X-Y). \tag{A.5}$$

From (A.2) and (A.5), we see that both $\widehat{\theta}_{\mathrm{GS}}$ and $\widehat{\theta}_{\mathrm{eo}}$ shrink the same unbiased estimator X to the same biased estimator Y but with different weights.

Lemma A.1. *When $k \geq 4$, the averaging estimator $\widehat{\theta}_{\mathrm{eo}}$ defined in* (A.5) *satisfies*

$$\mathbb{E}\left[\|\widehat{\theta}_{\mathrm{eo}} - \theta\|^2 - \|\widehat{\theta}_1 - \theta\|^2\right] < 0 \tag{A.6}$$

for any $\theta \in \mathbb{R}^k$, any $d \in \mathbb{R}^k$ and any $\sigma^2 > 0$.

The inequality (A.6) shows that the risk of the averaging GMM estimator is strictly smaller than that of the conservative GMM estimator if $k \geq 4$, for any $\theta \in \mathbb{R}^k$, any $d \in \mathbb{R}^k$ and any $\sigma^2 > 0$, and hence it uniformly dominates the MLE of θ. The condition on k for the uniform dominance result of our averaging estimator is slightly stronger than the condition for Green and Strawderman (1991)'s estimator. The proof of Lemma A.1 is given in Section E of the Supplemental Appendix. It is different from the proof for that in Green and Strawderman (1991) and Judge and Mittelhammer (2004) because the two averaging estimators are different. This proof is analogous to the proof of Theorem 5.2 for the general case. Thus we put it in the Appendix in the Online Supplemental Material of the paper (Cheng, Liao, and Shi (2019)).

Appendix B: Proof of results in Section 4 and Section 5

In Lemma 3.1, define

$$\begin{aligned} c_\rho &\equiv \min_{F^* \in \mathcal{F}^*} \{\rho_{\min}(\Gamma_{xz_1}\Gamma_{z_1x}), \rho_{\min}(\Psi)\}, \\ C_\rho &\equiv \max_{F^* \in \mathcal{F}^*} \{\|\phi\|^2, \rho_{\max}(\Psi)\}, \\ C_\Delta &\equiv \sup_{\delta_0 \in \Delta_\delta} \|\delta_0\|^2. \end{aligned} \tag{B.1}$$

Let

$$C_{*,W} \equiv 2(d_\theta + r_2 + 1)C_\rho, \qquad c_{*,\rho} \equiv \min\{1, c_\rho^2\} \quad \text{and} \quad C_{*,\rho} \equiv C_{*,W}^2 \big(2 + C_\Delta^{1/2}\big)^2. \tag{B.2}$$

Then, in Lemma 3.1(iii), the constant ε is given by

$$\varepsilon = c_{*,\rho} C_{*,\rho}^{-1} C_\Delta^{-1}, \tag{B.3}$$

that is, we require the condition to hold on a set bounded away from 0 by ε. The details of the proofs are given in Section D of the Supplemental Appendix.

B.1 *Proof of the results in Section 4*

Let $\mu_n(g_2(W, \theta)) = n^{-1/2} \sum_{i=1}^n (g_2(W_i, \theta) - \mathbb{E}_{F_n}[g_2(W_i, \theta)])$. In the rest of the Appendix, we use C to denote a generic fixed positive finite constant which does not depend on any $F \in \mathcal{F}$ or n.

Lemma B.1. *Suppose that Assumption* 3.2(ii) *holds and* Θ *is compact. Then we have*:

(i) $\sup_{\theta \in \Theta} \|\overline{g}_2(\theta) - \mathbb{E}_{F_n}[g_2(W_i, \theta)]\| = o_p(1)$;

(ii) $\sup_{\theta \in \Theta} \|n^{-1} \sum_{i=1}^n g_2(W_i, \theta) g_2(W_i, \theta)' - \mathbb{E}_{F_n}[g_2(W_i, \theta) g_2(W_i, \theta)']\| = o_p(1)$;

(iii) $\sup_{\theta \in \Theta} \|n^{-1} \sum_{i=1}^n g_{2,\theta}(W_i, \theta) - \mathbb{E}_{F_n}[g_{2,\theta}(W_i, \theta)]\| = o_p(1)$;

(iv) $\mu_n(g_2(W, \theta))$ *is stochastic equicontinuous over* $\theta \in \Theta$;

(v) $\Omega_{2,F_n}^{-1/2} \mu_n(g_2(W, \theta_{F_n})) \to_D N(0_{r_2 \times 1}, I_{r_2})$.

Proof of Lemma B.1. See Lemmas 11.3–11.5 of Andrews and Cheng (2013). □

Define $M_{k,F}(\theta) = \mathbb{E}_F[g_k(W, \theta)]$, $G_{k,F}(\theta) = \mathbb{E}_F[g_{k,\theta}(W, \theta)]$ and $\Omega_{k,F}(\theta) = \mathrm{Var}_F[g_k(W, \theta)]$, for any $F \in \mathcal{F}$, for any $\theta \in \Theta$ and for $k = 1, 2$. The next lemma shows that $M_{2,F}(\cdot)$, $G_{2,F}(\cdot)$ and $\Omega_{2,F}(\cdot)$ are Lipschitz continuous uniformly over $F \in \mathcal{F}$.

Lemma B.2. *Under Assumptions* 3.2(i)–(ii), *for any* $F \in \mathcal{F}$ *and any* $\theta_1, \theta_2 \in \Theta$, *we have*:

(i) $\|M_{2,F}(\theta_1) - M_{2,F}(\theta_2)\| \le C\|\theta_1 - \theta_2\|$;

(ii) $\|G_{2,F}(\theta_1) - G_{2,F}(\theta_2)\| \le C\|\theta_1 - \theta_2\|$;

(iii) $\|\Omega_{2,F}(\theta_1) - \Omega_{2,F}(\theta_2)\| \le C\|\theta_1 - \theta_2\|$.

Proof of Lemma B.2 is included in Section E of the Supplemental Appendix.

Lemma B.3. *Suppose that Assumptions* 3.1(i)–(ii) *and* 3.2(i)–(ii) *hold. Then for any sequence of DGPs* $\{F_n\}$*, we have*

$$\widetilde{\theta}_1 - \theta_{F_n} = o_p(1) \quad \text{and} \quad \overline{\Omega}_2 = \Omega_{2,F_n} + o_p(1), \tag{B.4}$$

where $\widetilde{\theta}_1$ *is a preliminary estimator defined as*

$$\widetilde{\theta}_1 = \arg\min_{\theta\in\Theta} \overline{g}_1(\theta)'\overline{g}_1(\theta) \tag{B.5}$$

and $\overline{\Omega}_2$ *is defined in* (E.13) *of the Supplemental Appendix.*

Proof of Lemma B.3 is included in Section E of the Supplemental Appendix.

Lemma B.4. *Suppose that Assumptions* 3.1(i)–(ii) *and* 3.2 *hold. Then for any sequence of DGPs* $\{F_n\}$*, we have*

$$n^{1/2}(\widehat{\theta}_1 - \theta_{F_n}) = \Gamma_{1,F_n}\mu_n\big(g_1(W, \theta_{F_n})\big) + o_p(1), \tag{B.6}$$

where $\Gamma_{1,F_n}\mu_n(g_1(W, \theta_{F_n})) \equiv -(G'_{1,F_n}\Omega^{-1}_{1,F_n}G_{1,F_n})^{-1}G'_{1,F_n}\Omega^{-1}_{1,F_n} = O_p(1)$.

Proof of Lemma B.4 is included in Section E of the Supplemental Appendix.

Lemma B.5. *Suppose that Assumptions* 3.1(iii) *and* 3.2(i)–(iii) *hold. Then for any sequence of DGPs* $\{F_n\}$*, we have*

$$\widehat{\theta}_2 - \theta^*_{F_n} = o_p(1). \tag{B.7}$$

Proof of Lemma B.5 is included in Section E of the Supplemental Appendix.

Lemma B.6. *Suppose that Assumptions* 3.1(i)–(ii) *and* 3.2(i)–(iii) *hold. Consider any sequence of DGPs* $\{F_n\}$ *such that* $\delta_{F_n} = o(1)$*. Then we have*

$$\widehat{\theta}_2 - \theta_{F_n} = o_p(1). \tag{B.8}$$

If we further have Assumption 3.2(iv)*, then*

$$n^{1/2}(\widehat{\theta}_2 - \theta_{F_n}) = \big(\Gamma_{2,F_n} + o_p(1)\big)\big\{\mu_n\big(g_2(W, \theta_{F_n})\big) + n^{1/2}\delta_{2,F_n}\big\} + o_p(1), \tag{B.9}$$

where $\Gamma_{2,F_n} = -(G'_{2,F_n}\Omega^{-1}_{2,F_n}G_{2,F_n})^{-1}G'_{2,F_n}\Omega^{-1}_{2,F_n}$ *and* $\delta_{2,F_n} = (\mathbf{0}_{1\times r_1}, \delta'_{F_n})'$.

Proof of Lemma B.6 is included in Section E of the Supplemental Appendix.

Lemma B.7. *Under Assumptions* 3.2(ii) *and* 3.3(ii)*, for any sequence of DGPs* $\{F_{p_n}\}$ *with* $F_{p_n} \in \mathcal{F}$ *where* $\{p_n\}$ *is a subsequence of* $\{n\}$*, there is a subsequence* $\{p_n^*\}$ *of* $\{p_n\}$ *such that* $v_{F_{p_n^*}}(\theta_{F_{p_n^*}}) \to v_F(\theta_F)$ *as* $p_n^* \to \infty$*, where* $F \in \mathcal{F}$.

Proof of Lemma B.7. Recall that $\Lambda = \{v_F : F \in \mathcal{F}\}$. By Assumptions 3.2(ii) and 3.3(ii), Λ is compact. Hence for any sequence $\{v_{F_{p_n}}(\theta_{F_{p_n}})\}$ in Λ, it has a convergent subsequence $\{v_{F_{p_n^*}}(\theta_{F_{p_n^*}})\}$ such that $v_{F_{p_n^*}}(\theta_{F_{p_n^*}}) \to v_F(\theta_F)$ as $p_n^* \to \infty$, where $F \in \mathcal{F}$. □

Lemma B.8. *Suppose that Assumptions* 3.1(i)–(ii) *and* 3.2 *hold. Consider any sequence of DGPs* $\{F_n\}$ *such that* $\overline{v}_{F_n} \to \overline{v}_F$ *for some* $F \in \mathcal{F}$, *and* $n^{1/2}\delta_{F_n} \to d$ *for* $d \in \mathbb{R}^{r^*}$. *Then*

$$\begin{pmatrix} n^{1/2}(\widehat{\theta}_1 - \theta_{F_n}) \\ n^{1/2}(\widehat{\theta}_2 - \theta_{F_n}) \end{pmatrix} \to_D \begin{pmatrix} \xi_{1,F} \\ \xi_{2,F} \end{pmatrix} \equiv \begin{pmatrix} \Gamma_{1,F}\mathcal{Z}_{1,F} \\ \Gamma_{2,F}(\mathcal{Z}_{2,F} + d_0) \end{pmatrix},$$

where $d_0 = (\mathbf{0}_{1\times r_1}, d')'$.

Proof of Lemma B.8. In the proof, we use

$$G_{2,F_n} \to G_{2,F} \quad \text{and} \quad \Omega_{2,F_n} \to \Omega_{2,F} \tag{B.10}$$

for some $F \in \mathcal{F}$, which is assumed in the lemma. Under Assumptions 3.1(i)–(ii) and 3.2, for the sequence of DGPs $\{F_n\}$ considered in the lemma, we can apply Lemma B.4 and Lemma B.6 to deduce that

$$\begin{pmatrix} n^{1/2}(\widehat{\theta}_1 - \theta_{F_n}) \\ n^{1/2}(\widehat{\theta}_2 - \theta_{F_n}) \end{pmatrix} = \begin{pmatrix} \Gamma_{1,F_n}\mu_n\big(g_1(W, \theta_{F_n})\big) \\ \big(\Gamma_{2,F_n} + o_p(1)\big)\big\{\mu_n\big(g_2(W, \theta_{F_n})\big) + n^{1/2}\delta_{2,F_n}\big\} \end{pmatrix} + o_p(1), \tag{B.11}$$

where $\delta_{2,F_n} = (\mathbf{0}_{1\times r_1}, \delta'_{F_n})'$. By (B.10) and Assumption 3.2, we have

$$\Gamma_{1,F_n} = \Gamma_{1,F} + o(1) \quad \text{and} \quad \Gamma_{2,F_n} = \Gamma_{2,F} + o(1), \tag{B.12}$$

where $\Gamma_{k,F} = -(G'_{k,F}\Omega^{-1}_{k,F}G_{k,F})^{-1}G'_{k,F}\Omega^{-1}_{k,F}$ for $k = 1, 2$. Collecting the results in Lemmas B.1(v), (B.11) and (B.12), and then applying the continuous mapping theorem (CMT), we have

$$\begin{pmatrix} n^{1/2}(\widehat{\theta}_1 - \theta_{F_n}) \\ n^{1/2}(\widehat{\theta}_2 - \theta_{F_n}) \end{pmatrix} \to_D \begin{pmatrix} \Gamma^*_{1,F} \\ \Gamma_{2,F} \end{pmatrix} (\mathcal{Z}_{2,F} + d_0), \tag{B.13}$$

where $\mathcal{Z}_{2,F} \sim N(\mathbf{0}_{r_2\times 1}, \Omega_{2,F})$, $\Gamma^*_{1,F} = (\Gamma_{1,F}, \mathbf{0}_{d_\theta\times r^*})$ and $d_0 = (\mathbf{0}_{1\times r_1}, d')'$. The claimed result follows from (B.13) and the definitions of $\Gamma^*_{1,F}$ and $\mathcal{Z}_{2,F}$. □

Proof of Lemma 4.1. The claimed result in Part (a) has been proved in Lemma B.8.

We next consider the case that $n^{1/2}\delta_{F_n} \to d$ with $\|d\| = \infty$. Note that the results in (B.6) and (B.12) do not depend on $\|d\| < \infty$ or $\|d\| = \infty$. Using (B.6), (B.12), Lemma B.1(v), and the CMT, we have

$$n^{1/2}(\widehat{\theta}_1 - \theta_{F_n}) \to_D \Gamma_{1,F}\mathcal{Z}_{1,F}. \tag{B.14}$$

To study the properties of $\widehat{\theta}_2$, we have to consider two separate scenarios: (1) $\delta_{F_n} = o(1)$; and (2) $\|\delta_{F_n}\| > c_\delta$ for some $c_\delta > 0$. In scenario (1), Assumption 3.2, Lemma B.1(v), and Lemma B.6 imply that

$$n^{1/2}(\widehat{\theta}_2 - \theta_{F_n}) = \big(\Gamma_{2,F_n} + o_p(1)\big)n^{1/2}\delta_{F_n} + O_p(1). \tag{B.15}$$

By Assumption 3.1(iv) and $\|n^{1/2}\delta_{F_n}\| \to \infty$,

$$n\delta'_{F_n}\Gamma'_{2,F_n}\Gamma_{2,F_n}\delta_{F_n} \geq C^{-2}n\delta'_{F_n}\delta_{F_n} \to \infty, \tag{B.16}$$

which together with (B.15) implies that $\|n^{1/2}(\widehat{\theta}_2 - \theta_{F_n})\| \to_p \infty$.

Finally, we consider the scenario (2) where $\|\delta_{F_n}\| > c_\delta$. By Assumption 3.1(iv),

$$\big\|G'_{2,F_n}\Omega^{-1}_{2,F_n}\delta_{F_n}\big\| > C^{-1}\|\delta_{F_n}\| > c_\delta C^{-1} \tag{B.17}$$

for any n. As $\theta^*_{F_n}$ is the minimizer of $Q_{F_n}(\theta)$, it has the following first-order condition:

$$0_{d_\theta \times 1} = G_{2,F_n}\big(\theta^*_{F_n}\big)'\Omega^{-1}_{2,F_n}M_{2,F_n}\big(\theta^*_{F_n}\big), \tag{B.18}$$

which implies that

$$\begin{aligned} G'_{2,F_n}\Omega^{-1}_{2,F_n}\delta_{F_n} &= G_{2,F_n}(\theta_{F_n})'\Omega^{-1}_{2,F_n}M_{2,F_n}(\theta_{F_n}) - G_{2,F_n}\big(\theta^*_{F_n}\big)'\Omega^{-1}_{2,F_n}M_{2,F_n}\big(\theta^*_{F_n}\big) \\ &= \big[G_{2,F_n}(\theta_{F_n}) - G_{2,F_n}\big(\theta^*_{F_n}\big)\big]'\Omega^{-1}_{2,F_n}M_{2,F_n}(\theta_{F_n}) \\ &\quad + G_{2,F_n}\big(\theta^*_{F_n}\big)'\Omega^{-1}_{2,F_n}\big[M_{2,F_n}(\theta_{F_n}) - M_{2,F_n}\big(\theta^*_{F_n}\big)\big]. \end{aligned} \tag{B.19}$$

By Lemma B.2, the Cauchy–Schwarz inequality and Assumption 3.2(ii)–(iii), we have

$$\begin{aligned} &\big\|\big[G_{2,F_n}(\theta_{F_n}) - G_{2,F_n}\big(\theta^*_{F_n}\big)\big]'\Omega^{-1}_{2,F_n}M_{2,F_n}(\theta_{F_n})\big\| \\ &\quad \leq \big\|G_{2,F_n}(\theta_{F_n}) - G_{2,F_n}\big(\theta^*_{F_n}\big)\big\|\big\|\Omega^{-1}_{2,F_n}M_{2,F_n}(\theta_{F_n})\big\| \leq C\big\|\theta_{F_n} - \theta^*_{F_n}\big\|, \end{aligned} \tag{B.20}$$

where C is a fixed constant. Similarly, we have

$$\begin{aligned} &\big\|G_{2,F_n}\big(\theta^*_{F_n}\big)'\Omega^{-1}_{2,F_n}\big[M_{2,F_n}(\theta_{F_n}) - M_{2,F_n}\big(\theta^*_{F_n}\big)\big]\big\| \\ &\quad \leq \big\|M_{2,F_n}(\theta_{F_n}) - M_{2,F_n}\big(\theta^*_{F_n}\big)\big\|\|\Omega^{-1}_{2,F_n}G_{2,F_n}\theta^*_{F_n})\| \leq C\big\|\theta_{F_n} - \theta^*_{F_n}\big\|. \end{aligned} \tag{B.21}$$

Combining the results in (B.19), (B.20), and (B.21), and using the triangle inequality, we have

$$\big\|\theta_{F_n} - \theta^*_{F_n}\big\| \geq c_\delta C \tag{B.22}$$

for some fixed constant C. Using $\widehat{\theta}_2 = \theta^*_{F_n} + o_p(1)$ (which is proved in Lemma B.5) and the triangle inequality, we obtain

$$\|\widehat{\theta}_2 - \theta_{F_n}\| \geq \big|\big\|\widehat{\theta}_2 - \theta^*_{F_n}\big\| - \big\|\theta^*_{F_n} - \theta_{F_n}\big\|\big| = \big\|\theta^*_{F_n} - \theta_{F_n}\big\|\big(1 + o_p(1)\big), \tag{B.23}$$

which together with (B.22) implies that $n^{1/2}\|\widehat{\theta}_2 - \theta_{F_n}\| \to_p \infty$. This completes the proof. □

Lemma B.9. (a) $\Gamma^*_{1,F}d_0 = \mathbf{0}_{d_\theta \times 1}$; (b) $\Gamma^*_{1,F}\Omega_{2,F}\Gamma^{*\prime}_{1,F} = \Sigma_{1,F}$; (c) $\Gamma^*_{1,F}\Omega_{2,F}\Gamma'_{2,F} = \Sigma_{2,F}$; (d) $\Gamma_{2,F}\Omega_{2,F}\Gamma'_{2,F} = \Sigma_{2,F}$.

Proof of Lemma B.9 is included in Section E of the Supplemental Appendix.

B.2 *Proof of the results in Section* 5

We first present some generic results on the bounds of asymptotic risk difference between two estimators under some high-level conditions. Then we apply these generic results to the two specific estimators we consider in this paper: $\widehat{\theta}_{\text{eo}}$ and $\widehat{\theta}_1$. The proof uses the subsequence techniques used to show the asymptotic size of a test in Andrews, Cheng, and Guggenberger (2011) but we adapt the proof and notation to the current setup and extend results from test to estimators.

Recall that $h_{F,d} = (d', \text{vec}(G_{2,F})', \text{vech}(\Omega_{2,F})')$ and $\overline{v}_F = (\text{vec}(G_{2,F})', \text{vech}(\Omega_{2,F})')$ for any $F \in \mathcal{F}$ and any $d \in \mathbb{R}^{r^*}_\infty$. We have defined

$$H = \{h_{F,d} : d \in \mathbb{R}^{r^*} \text{ and } F \in \mathcal{F} \text{ with } \delta_F = 0_{r^* \times 1}\}, \tag{B.24}$$

where δ_F is defined by (1.5) for a given F. Define

$$H^*_\infty = \{h_{F,d} : d \in \mathbb{R}^{r^*}_\infty \text{ with } \|d\| = \infty \text{ and } F \in \mathcal{F}\}. \tag{B.25}$$

Let $d_h = r^* + d_\theta r_2 + (r_2 + 1)r_2/2$. It is clear that $h_{F,d}$ is a d_h-dimensional vector.

Condition B.1. (i) *For any sequence of DGPs* $\{F_{p_n}\}$ *with* $F_{p_n} \in \mathcal{F}$ *where* $\{p_n\}$ *is a subsequence of* $\{n\}$*, there exists a subsequence* $\{p_n^*\}$ *of* $\{p_n\}$ *and some* $F \in \mathcal{F}$ *such that* $v_{F_{p_n^*}} \to v_F$ *as* $p_n^* \to \infty$:

(ii) $M_{1,F}(\theta) = 0_{r_1 \times 1}$ *has a unique solution at* $\theta_F \in \Theta$ *for any* $F \in \mathcal{F}$;

(iii) $M_{2,F}(\cdot)$ *is uniform equicontinuous over* $F \in \mathcal{F}$;

(iv) *for any subsequence* $\{p_n\}$ *of* $\{n\}_{n\in\mathbb{N}}$*, if* $(p_n)^{1/2}\delta_{F_{p_n}} \to d$ *for* $d \in \mathbb{R}^{r^*}_\infty$ *and* $v_{F_{p_n}} \to v_F$*, then*

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n}})\big] = R_\zeta(h_{F,d}) \quad \textit{and} \quad \lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widetilde{\theta}, \theta_{F_{p_n}})\big] = \widetilde{R}_\zeta(h_{F,d}),$$

where $R_\zeta(h_{F,d})$ *and* $\widetilde{R}_\zeta(h_{F,d})$ *are some nonnegative functions that are bounded from above by* ζ *for any* $F \in \mathcal{F}$ *and any* $d \in \mathbb{R}^{r^*}_\infty$;

(v) *for any* $F \in \mathcal{F}$ *with* $\delta_F = 0_{r^* \times 1}$*, there exists a constant* $\varepsilon_F > 0$ *such that for any* $\widetilde{\delta} \in \mathbb{R}^{r^*}$ *with* $0 \leq \|\widetilde{\delta}\| < \varepsilon_F$*, there is* $\widetilde{F} \in \mathcal{F}$ *with* $\delta_{\widetilde{F}} = \widetilde{\delta}$ *and* $\|\overline{v}_F - \overline{v}_{\widetilde{F}}\| \leq C\|\widetilde{\delta}\|^\kappa$ *for some* $\kappa > 0$;

(vi) *for any* $h_{F,d} \in H^*_\infty$ *and* $h_{F,\widetilde{d}} \in H^*_\infty$*, we have*

$$R_\zeta(h_{F,d}) = R_\zeta(h_{F,\widetilde{d}}) \quad \textit{and} \quad \widetilde{R}_\zeta(h_{F,d}) = \widetilde{R}_\zeta(h_{F,\widetilde{d}})$$

for any $\zeta > 0$.

Condition B.1(i) requires that for any sequence of $\{v_{F_{p_n}}\}$, it has a convergent subsequence $\{v_{F_{p_n^*}}\}$ with limit being v_F for some $F \in \mathcal{F}$. This condition is verified under Assumptions 3.2(ii) and 3.3(ii) in Lemma B.7. Condition B.1(ii) is the unique identification condition of θ_F which holds under Assumptions 3.1(i)–(ii). Condition B.1(iii) holds under Assumption 3.2(ii) by Lemma B.2. Condition B.1(iv) is a key assumption to derive an explicit upper bound of asymptotic risk. This condition can be verified by using Lemma

4.1 as we shall show in the proof of Theorem 5.1. Condition B.1(v) enables us to show that the upper bound we derived for the asymptotic risk is also a lower bound. This condition is assumed in Assumption 3.3(i). Condition B.1(vi), in our context, requires that the asymptotic (truncated) risk of $\widehat{\theta}$ (or $\widetilde{\theta}$) under the subsequences of DGPs $\{F_{p_n}\}$ satisfying the restrictions in Condition B.1(iv) are identical whenever $(p_n)^{1/2}\delta_{F_{p_n}} \to d$ with $\|d\| = \infty$. Condition B.1 is verified in the proof of Theorem 5.1 below.

Lemma B.10. *Under Conditions* B.1(i)–B.1(iv), *we have*

$$\mathrm{Asy}R_\zeta(\widehat{\theta}) \le \max\Big\{\sup_{h\in H} R_\zeta(h), \sup_{h\in H^*_\infty} R_\zeta(h)\Big\}, \tag{B.26}$$

where $\mathrm{Asy}R_\zeta(\widehat{\theta}) \equiv \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{E}_F[\ell_\zeta(\widehat{\theta}, \theta_F)]$.

Proof of Lemma B.10. Let $\{F_n\}$ be a sequence such that

$$\limsup_{n\to\infty}\mathbb{E}_{F_n}\big[\ell_\zeta(\widehat{\theta}, \theta_{F_n})\big] = \limsup_{n\to\infty}\Big(\sup_{F\in\mathcal{F}} \mathbb{E}_F\big[\ell_\zeta(\widehat{\theta}, \theta_F)\big]\Big) \equiv \mathrm{Asy}R_\zeta(\widehat{\theta}). \tag{B.27}$$

Such a sequence always exists by the definition of supremum. The sequence $\{\mathbb{E}_{F_n}[\ell_\zeta(\widehat{\theta}, \theta_{F_n})]: n \ge 1\}$ may not converge. However, by the definition of limsup, there exists a subsequence of $\{n\}_{n\in\mathbb{N}}$, say $\{p_n\}$, such that $\{\mathbb{E}_{F_{p_n}}[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n}})]: n \ge 1\}$ converges and

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n}})\big] = \mathrm{Asy}R_\zeta(\widehat{\theta}). \tag{B.28}$$

Below we show that for any subsequence $\{p_n\}$ of $\{n\}_{n\in\mathbb{N}}$ such that $\{\mathbb{E}_{F_{p_n}}[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n}})]: n \ge 1\}$ is convergent, there exists a subsequence $\{p_n^*\}$ of $\{p_n\}$ such that

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n^*}}\big[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n^*}})\big] = R_\zeta(h) \quad \text{for some } h \in H \text{ or } H^*_\infty. \tag{B.29}$$

Because $\lim_{n\to\infty} \mathbb{E}_{F_{p_n^*}}[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n^*}})] = \lim_{n\to\infty} \mathbb{E}_{F_{p_n}}[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n}})]$, which combined with (B.28) and (B.29) implies that

$$\mathrm{Asy}R_\zeta(\widehat{\theta}) = R_\zeta(h) \quad \text{for some } h \in H \text{ or } H^*_\infty. \tag{B.30}$$

The desired result in (B.26) follows immediately by (B.30).

To show that there exists a subsequence $\{p_n^*\}$ of $\{p_n\}$ such that (B.29) holds, it suffices to show that for any sequence $\{F_n\}$ and any subsequence $\{p_n\}$ of $\{n\}_{n\in\mathbb{N}}$, there exists a subsequence $\{p_n^*\}$ of $\{p_n\}$ for which we have

$$(p_n^*)^{1/2}\delta_{F_{p_n^*}} \to d \quad \text{for } d \in \mathbb{R}^{r^*}_\infty \text{ and } v_{F_{p_n^*}} \to v_F \tag{B.31}$$

for some $F \in \mathcal{F}$. If (B.31) holds, then we can use Condition B.1(iv) to deduce that

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n^*}}\big[\ell_\zeta(\widehat{\theta}, \theta_{F_{p_n^*}})\big] = R_\zeta(h_{F,d}) \tag{B.32}$$

for the sequence of DGPs $\{F_{p_n^*}\}$ that satisfies (B.31). As $d \in \mathbb{R}^{r^*}_\infty$, we have either $\|d\| < \infty$ or $\|d\| = \infty$. In the first case, $\|d\| < \infty$ together with $(p_n^*)^{1/2}\delta_{F_{p_n^*}} \to d$ and $\delta_{F_{p_n^*}} \to \delta_F$ (which is implied by $v_{F_{p_n^*}} \to v_F$) implies that $\delta_F = 0_{r^*\times 1}$, which implies that $h_{F,d} \in H$ by the definition of H. In the second case, $h_{F,d} \in H_\infty^*$ by the definition of H_∞^*. We have proved that $h_{F,d}$ in (B.32) belongs either to H or H_∞^* which together with (B.32) proves (B.29).

Finally, we show that for any sequence $\{F_n\}$ and any subsequence $\{p_n\}$ of $\{n\}_{n\in\mathbb{N}}$, there exists a subsequence $\{p_n^*\}$ of $\{p_n\}$ for which (B.31) holds. Let $\delta_{p_n,j}$ denote the jth component of δ_{p_n} and $p_{1,n} = p_n$ for any $n \geq 1$. For $j = 1$, either (a) $\limsup_{n\to\infty} |p_{j,n}^{1/2} \times \delta_{p_{j,n},j}| < \infty$; or (b) $\limsup_{n\to\infty} |p_{j,n}^{1/2}\delta_{p_{j,n},j}| = \infty$. If (a) holds, then for some subsequence $\{p_{j+1,n}\}$ of $\{p_{j,n}\}$, $p_{j+1,n}^{1/2}\delta_{p_{j+1,n},j} \to d_j$ for some $d_j \in \mathbb{R}$. If (b) holds, then for some subsequence $\{p_{j+1,n}\}$ of $\{p_{j,n}\}$, $p_{j+1,n}^{1/2}\delta_{p_{j+1,n},j} \to \infty$ or $-\infty$. As r^* is a fixed positive integer, we can apply the same arguments successively for $j = 1, \dots, r^*$ to obtain a subsequence $\{p_{r^*,n}\}$ of $\{p_n\}$ such that $(p_{r^*,n})^{1/2}\delta_{p_{r^*,n}} \to d \in \mathbb{R}_\infty^{r^*}$. By Condition B.1(i), we know that there exists a subsequence $\{p_n^*\}$ of $\{p_{r^*,n}\}$ such that $v_{p_n^*} \to v_F$ for some $F \in \mathcal{F}$, which completes the proof of (B.31). □

Lemma B.11. *Suppose that Condition* B.1(v) *holds. Then* (i) *for any $h_{F,d} \in H$, there exists a sequence of DGPs $\{F_n\}$ with $F_n \in \mathcal{F}$ such that*

$$n^{1/2}\delta_{F_n} \to d, \qquad G_{2,F_n} \to G_{2,F} \quad \textit{and} \quad \Omega_{2,F_n} \to \Omega_{2,F}; \tag{B.33}$$

(ii) *for any $h_{F,d} \in H_\infty^*$, there exists a sequence of DGPs $\{F_n\}$ with $F_n \in \mathcal{F}$ such that*

$$\|n^{1/2}\delta_{F_n}\| \to \infty, \qquad G_{2,F_n} \to G_{2,F}, \qquad \Omega_{2,F_n} \to \Omega_{2,F} \quad \textit{and} \quad \delta_{F_n} \to \delta_F. \tag{B.34}$$

Proof of Lemma B.11. (i) By the definition of H, we have $\delta_F = 0_{r^*\times 1}$ for any F such that $h_{F,d} \in H$. Let N_{ε_F} be the smallest n such that $\|d\|n^{-1/2} < \varepsilon_F$. By Condition B.1(v), for any $n \geq N_{\varepsilon_F}$ we can find a DGP F_n such that

$$\delta_{F_n} = n^{-1/2}d \quad \text{and} \quad \|\overline{v}_{F_n} - \overline{v}_F\| \leq n^{-\kappa/2}C\|d\|^\kappa. \tag{B.35}$$

For any $n < N_{\varepsilon_F}$ such that $\|d\|n^{-1/2} \geq \varepsilon_F$, we let $F_n = F$. The desired properties in (B.33) holds under the constructed sequence of DGPs $\{F_n\}$ by (B.35), because C is a fixed constant and $\kappa > 0$.

(ii) For any $h_{F,d} \in H_\infty^*$, we have either $\delta_F = 0_{r^*\times 1}$ or $\|\delta_F\| > 0$. We first consider the case that $\delta_F = 0_{r^*\times 1}$. Let $1_{r^*\times 1}$ denote the $r^* \times 1$ vector of ones. Let N_{ε_F} be the smallest n such that $n^{-1/4}(r^*)^{1/2} < \varepsilon_F$. By Condition B.1(v), for any $n \geq N_{\varepsilon_F}$ we can find a DGP F_n such that

$$\delta_{F_n} = n^{-1/4}1_{r^*\times 1} \quad \text{and} \quad \|\overline{v}_{F_{p_n}} - \overline{v}_F\| \leq Cn^{-\kappa/4}\bigl(r^*\bigr)^{\kappa/2}. \tag{B.36}$$

For any $n < N_{\varepsilon_F}$ such that $n^{-1/4}(r^*)^{1/2} \geq \varepsilon_F$, we let $F_n = F$. The desired properties in (B.34) holds under the constructed sequence of DGPs $\{F_n\}$ by (B.36), because C is a fixed constant and $\kappa > 0$. When $\|\delta_F\| > 0$, we define a trivial sequence of DGPs $\{F_n\}$ as $F_n = F$ for any n. It is clear that (B.34) holds trivially in this case. □

Lemma B.12. *Under Condition* B.1, *we have*

$$\text{Asy}R_\zeta(\widehat{\theta}) = \max\Bigl\{\sup_{h\in H} R_\zeta(h), \sup_{h\in H_\infty^*} R_\zeta(h)\Bigr\}. \tag{B.37}$$

Proof of Lemma B.12. In view of the upper bound in (B.26) in Lemma B.10, it is sufficient to show that

$$\mathrm{Asy}R_{\zeta}(\widehat{\theta}) \geq \max\Big\{\sup_{h\in H} R_{\zeta}(h), \sup_{h\in H_{\infty}^{*}} R_{\zeta}(h)\Big\}. \tag{B.38}$$

First, we note that for any $h_{d,F} = (d', \mathrm{vec}(G_{2,F})', \mathrm{vech}(\Omega_{2,F})') \in H$, there exists a sequence $\{F_n \in \mathcal{F} : n \geq 1\}$ such that

$$n^{1/2}\delta_{F_n} \to d \in \mathbb{R}^{r^{*}} \quad \text{and} \quad v_{F_n} \to v_F \tag{B.39}$$

by Lemma B.11(i). The sequence $\mathbb{E}_{F_n}[\ell_{\zeta}(\widehat{\theta}, \theta_{F_n})]$ may not be convergent, but there exists a subsequence $\{p_n\}$ of n such that $\mathbb{E}_{F_{p_n}}[\ell_{\zeta}(\widehat{\theta}, \theta_{F_{p_n}})]$ is convergent and

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\big[\ell(\widehat{\theta}, \theta_{F_{p_n}})\big] = \limsup_{n\to\infty} \mathbb{E}_{F_n}\big[\ell(\widehat{\theta}, \theta_{F_n})\big]. \tag{B.40}$$

As $\{p_n\}$ is a subsequence of $\{n\}_{n\in\mathbb{N}}$, by (B.39)

$$(p_n)^{1/2}\delta_{F_{p_n}} \to d \in \mathbb{R}^{r^{*}} \quad \text{and} \quad v_{F_{p_n}} \to v_F. \tag{B.41}$$

By Condition B.1(iv), we have that

$$\lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\big[\ell(\widehat{\theta}, \theta_{F_{p_n}})\big] = R_{\zeta}(h_{F,d}), \tag{B.42}$$

which combined with (B.40) and the definition of $\mathrm{Asy}R_{\zeta}(\widehat{\theta})$ gives

$$\mathrm{Asy}R_{\zeta}(\widehat{\theta}) = \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{E}_F\big[\ell_{\zeta}(\widehat{\theta}, \theta_F)\big] \geq \limsup_{n\to\infty} \mathbb{E}_{F_n}\big[\ell(\widehat{\theta}, \theta_{F_n})\big] = R_{\zeta}(h_{F,d}). \tag{B.43}$$

Second, consider any $h_{d,F} = (d', \mathrm{vec}(G_{2,F})', \mathrm{vech}(\Omega_{2,F})') \in H_{\infty}^{*}$. By Lemma B.11(ii), there exists a sequence of DGPs $\{F_n\}$ such that

$$\|n^{1/2}\delta_{F_n}\| \to \infty \quad \text{and} \quad v_{F_n} \to v_F. \tag{B.44}$$

Using the same arguments in proving (B.40) to (B.42), we can show that for some subsequence $\{p_n\}$ of $\{n\}_{n\in\mathbb{N}}$,

$$|p_n^{1/2}\delta_{F_{p_n}}\| \to \infty \quad \text{and} \quad v_{p_n} \to v_F \tag{B.45}$$

and

$$\limsup_{n\to\infty} \mathbb{E}_{F_n}\big[\ell(\widehat{\theta}, \theta_{F_n})\big] = \lim_{n\to\infty} \mathbb{E}_{F_{p_n}}\big[\ell(\widehat{\theta}, \theta_{F_{p_n}})\big] = R_{\zeta}(h_{F,d}). \tag{B.46}$$

for $\|d\| = \infty$ by Conditions B.1(vi). By the definition of $\mathrm{Asy}R_{\zeta}(\widehat{\theta})$ and (B.46),

$$\mathrm{Asy}R_{\zeta}(\widehat{\theta}) = \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{E}_F\big[\ell_{\zeta}(\widehat{\theta}, \theta_F)\big] \geq \limsup_{n\to\infty} \mathbb{E}_{F_n}\big[\ell(\widehat{\theta}, \theta_{F_n})\big] = R_{\zeta}(h_{F,d}). \tag{B.47}$$

Combining the results in (B.43) and (B.47), we immediately get (B.37). □

Lemma B.13. *Under Conditions* B.1(i)–B.1(iv), *the upper and lower bounds of the asymptotic risk difference between $\widehat{\theta}$ and $\widetilde{\theta}$ satisfy*

$$\mathrm{Asy}\overline{RD}(\widehat{\theta},\widetilde{\theta}) \le \lim_{\zeta\to\infty}\Bigl(\max\Bigl\{\sup_{h\in H}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr],\ \sup_{h\in H_\infty^*}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr]\Bigr\}\Bigr), \tag{B.48}$$

$$\mathrm{Asy}\underline{RD}(\widehat{\theta},\widetilde{\theta}) \ge \lim_{\zeta\to\infty}\Bigl(\min\Bigl\{\inf_{h\in H}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr],\ \inf_{h\in H_\infty^*}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr]\Bigr\}\Bigr), \tag{B.49}$$

where

$$\widetilde{R}_\zeta(h) \equiv \mathbb{E}\bigl[\min\{\xi_{1,F}'\Upsilon\xi_{1,F},\zeta\}\bigr] \quad \textit{and} \quad R_\zeta(h) \equiv \begin{cases}\mathbb{E}\bigl[\min\{\overline{\xi}_F'\Upsilon\overline{\xi}_F,\zeta\}\bigr], & \|d\|<\infty,\\ \mathbb{E}\bigl[\min\{\xi_{1,F}'\Upsilon\xi_{1,F},\zeta\}\bigr], & \|d\|=\infty,\end{cases}$$

for any $h\in H\cup H_\infty^*$.

Proof of Lemma B.13. Define

$$\overline{R}_\zeta\bigl(H,H_\infty^*\bigr) \equiv \max\Bigl\{\sup_{h\in H}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr],\ \sup_{h\in H_\infty^*}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr]\Bigr\}, \tag{B.50}$$

$$\underline{R}_\zeta\bigl(H,H_\infty^*\bigr) \equiv \min\Bigl\{\inf_{h\in H}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr],\ \inf_{h\in H_\infty^*}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr]\Bigr\}. \tag{B.51}$$

By the definition of $\mathrm{Asy}\overline{RD}(\widehat{\theta},\widetilde{\theta})$, to show (B.48) it is sufficient to show that for any $\zeta>0$

$$\limsup_{n\to\infty}\sup_{F\in\mathcal{F}}\mathbb{E}_F\bigl[\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)\bigr] \le \overline{R}_\zeta\bigl(H,H_\infty^*\bigr), \tag{B.52}$$

which can be proved using the same arguments in the proof of Lemma B.10 (but replacing $\ell_\zeta(\widehat{\theta},\theta_F)$ and $R_\zeta(h)$ by $\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)$ and $R_\zeta(h)-\widetilde{R}_\zeta(h)$, respectively). Similarly, by the definition of $\mathrm{Asy}\underline{RD}(\widehat{\theta},\widetilde{\theta})$, for (B.49) it is sufficient to show that for any $\zeta>0$,

$$\liminf_{n\to\infty}\inf_{F\in\mathcal{F}}\mathbb{E}_F\bigl[\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)\bigr] \ge \underline{R}_\zeta\bigl(H,H_\infty^*\bigr), \tag{B.53}$$

which can be proved using the same arguments in the proof of Lemma B.10 (but replacing $\limsup_n$, $\sup_{F\in\mathcal{F}}$, $\ell_\zeta(\widehat{\theta},\theta_F)$, and $R_\zeta(h)$ by $\liminf_n$, $\inf_{F\in\mathcal{F}}$, $\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)$, and $R_\zeta(h)-\widetilde{R}_\zeta(h)$, respectively). □

Lemma B.14. *Under Condition* B.1, *the upper and lower bounds of the asymptotic risk difference between $\widehat{\theta}$ and $\widetilde{\theta}$ have the following representations*:

$$\mathrm{Asy}\overline{RD}(\widehat{\theta},\widetilde{\theta}) = \lim_{\zeta\to\infty}\Bigl(\max\Bigl\{\sup_{h\in H}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr],\ \sup_{h\in H_\infty^*}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr]\Bigr\}\Bigr), \tag{B.54}$$

$$\mathrm{Asy}\underline{RD}(\widehat{\theta},\widetilde{\theta}) = \lim_{\zeta\to\infty}\Bigl(\min\Bigl\{\inf_{h\in H}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr],\ \inf_{h\in H_\infty^*}\bigl[R_\zeta(h)-\widetilde{R}_\zeta(h)\bigr]\Bigr\}\Bigr). \tag{B.55}$$

Proof of Lemma B.14. By Lemma B.13, it is sufficient to show that

$$\limsup_{n\to\infty}\sup_{F\in\mathcal{F}}\mathbb{E}_F\big[\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)\big]\geq\overline{R}_\zeta\big(H,H^*_\infty\big),\tag{B.56}$$

$$\liminf_{n\to\infty}\inf_{F\in\mathcal{F}}\mathbb{E}_F\big[\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)\big]\leq\underline{R}_\zeta\big(H,H^*_\infty\big),\tag{B.57}$$

for any $\zeta>0$. Equation (B.56) can be proved using the same arguments in the proof of Lemma B.12 by replacing $\ell_\zeta(\widehat{\theta},\theta_F)$ and $R_\zeta(h)$ by $\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)$ and $R_\zeta(h)-\widetilde{R}_\zeta(h)$, respectively. Similarly, (B.57) can be proved sing the same arguments in the proof of Lemma B.12 by replacing $\limsup_n$, $\sup_{F\in\mathcal{F}}$, $\ell_\zeta(\widehat{\theta},\theta_F)$ and $R_\zeta(h)$ by $\liminf_n$, $\inf_{F\in\mathcal{F}}$, $\ell_\zeta(\widehat{\theta},\theta_F)-\ell_\zeta(\widetilde{\theta},\theta_F)$ and $R_\zeta(h)-\widetilde{R}_\zeta(h)$, respectively. □

Lemma B.15. *Under Assumptions* 3.2(ii) *and* 3.2(iv), *we have*

$$\sup_{h\in H}\mathbb{E}\big[\big(\xi'_{1,F}\Upsilon\xi_{1,F}\big)^2\big]\leq C\quad\textit{and}\quad\sup_{h\in H}\mathbb{E}\big[\big(\overline{\xi}'_F\Upsilon\overline{\xi}_F\big)^2\big]\leq C.\tag{B.58}$$

Lemma B.16. *Let* $g_\zeta(h)\equiv\mathbb{E}[\min\{\overline{\xi}'_F\Upsilon\overline{\xi}_F,\zeta\}-\min\{\xi'_{1,F}\Upsilon\xi_{1,F},\zeta\}]$. *Under Assumptions* 3.2(ii) *and* 3.2(iv), *we have*

$$\lim_{\zeta\to\infty}\sup_{h\in H}\big[\big|g_\zeta(h)-g(h)\big|\big]=0,\tag{B.59}$$

where $\sup_{h\in H}[|g(h)|]\leq C$.

Proof of Theorem 5.1. The proof consists of two steps. The first step is to apply Lemma B.14 to show (B.60) and (B.61) below, and the second step is to apply Lemma B.16 to show (B.75) and (B.76) below.

In the first step, we apply Lemma B.14 with $\widehat{\theta}=\widehat{\theta}_{\text{eo}}$ and $\widetilde{\theta}=\widehat{\theta}_1$ to show that

$$\text{Asy}\overline{RD}(\widehat{\theta}_{\text{eo}},\widehat{\theta}_1)=\lim_{\zeta\to\infty}\max\Big\{\sup_{h\in H}\big[g_\zeta(h)\big],0\Big\}\quad\text{and}\tag{B.60}$$

$$\text{Asy}\underline{RD}(\widehat{\theta}_{\text{eo}},\widehat{\theta}_1)=\lim_{\zeta\to\infty}\min\Big\{\inf_{h\in H}\big[g_\zeta(h)\big],0\Big\}.\tag{B.61}$$

To prove (B.60) and (B.61), we now verify Condition B.1 under Assumptions 3.1–3.3. Condition B.1(i) is verified by Lemma B.7 under Assumptions 3.2(ii) and 3.3(ii). Condition B.1(ii) is implied by Assumptions 3.1(i) and 3.1(ii). Condition B.1(iii) is implied by Assumptions 3.2.(i)–(ii) as a result of Lemma B.2. Condition B.1(v) is assumed in Assumption 3.3(ii). We next verify Conditions B.1(iv) and B.1(vi).

Consider any sequence of DGPs $\{F_{p_n}\}$ with

$$(p_n)^{1/2}\delta_{F_{p_n}}\to d\quad\text{for } d\in\mathbb{R}^{r^*}_\infty\text{ and } v_{F_{p_n}}\to v_F\tag{B.62}$$

for some $F\in\mathcal{F}$, where $\{p_n\}$ is a subsequence of $\{n\}_{n\in\mathbb{N}}$. First, we consider the case that $d\in\mathbb{R}^{r^*}$. By Lemma 4.1(a) and 4.2(a),

$$(p_n)^{1/2}(\widehat{\theta}_1-\theta_{F_{p_n}})\to_D\xi_{1,F}\quad\text{and}\quad(p_n)^{1/2}(\widehat{\theta}_{\text{eo}}-\theta_{F_{p_n}})\to_D\overline{\xi}_F,\tag{B.63}$$

which combined with the continuous mapping theorem implies that

$$\ell(\widehat{\theta}_1, \theta_{F_{p_n}}) \to_D \xi_{1,F}' \Upsilon \xi_{1,F} \quad \text{and} \quad \ell(\widehat{\theta}_{\text{eo}}, \theta_{F_{p_n}}) \to_D \overline{\xi}_F' \Upsilon \overline{\xi}_F. \tag{B.64}$$

Since Υ is positive semidefinite, $\xi_{1,F}' \Upsilon \xi_{1,F}$ and $\overline{\xi}_F' \Upsilon \overline{\xi}_F$ are both nonnegative. The function $f_\zeta(x) = \min\{x, \zeta\}$ is a bounded continuous function for $x \geq 0$. By (B.64) and the Portmanteau lemma (see Lemma 2.2 in van der Vaart (1998)),

$$\begin{aligned} &\mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}_{\text{eo}}, \theta_{F_{p_n}})\big] \to \mathbb{E}\big[\min\big\{\overline{\xi}_F' \Upsilon \overline{\xi}_F, \zeta\big\}\big] \quad \text{and} \\ &\mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}_1, \theta_{F_{p_n}})\big] \to \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big]. \end{aligned} \tag{B.65}$$

Second, we consider the case that $\|d\| = \infty$. Then under Lemma 4.1(b) and 4.2(b),

$$(p_n)^{1/2}(\widehat{\theta}_1 - \theta_{F_{p_n}}) \to_D \xi_{1,F} \quad \text{and} \quad (p_n)^{1/2}(\widehat{\theta}_{\text{eo}} - \theta_{F_{p_n}}) \to_D \xi_{1,F}. \tag{B.66}$$

Using the same arguments in showing (B.65), we get

$$\begin{aligned} &\mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}_{\text{eo}}, \theta_{F_{p_n}})\big] \to \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big] \quad \text{and} \\ &\mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}_1, \theta_{F_{p_n}})\big] \to \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big]. \end{aligned} \tag{B.67}$$

Define

$$\begin{aligned} &\widetilde{R}_\zeta(h_{F,d}) = \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big] \quad \text{and} \\ &R_\zeta(h_{F,d}) = \begin{cases} \mathbb{E}\big[\min\big\{\overline{\xi}_F' \Upsilon \overline{\xi}_F, \zeta\big\}\big], & \|d\| < \infty, \\ \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big], & \|d\| = \infty. \end{cases} \end{aligned} \tag{B.68}$$

Collecting the results in (B.65) and (B.67), we deduce that under the sequence of DGPs $\{F_{p_n}\}$ satisfying (B.62),

$$\mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}_{\text{eo}}, \theta_{F_{p_n}})\big] \to R_\zeta(h_{F,d}) \quad \text{and} \quad \mathbb{E}_{F_{p_n}}\big[\ell_\zeta(\widehat{\theta}_1, \theta_{F_{p_n}})\big] \to \widetilde{R}_\zeta(h_{F,d}), \tag{B.69}$$

where $R_\zeta(h_{F,d})$ and $\widetilde{R}_\zeta(h_{F,d})$ are nonnegative and bounded from above by ζ for any $d \in \mathbb{R}_\infty^{r^*}$ and any $F \in \mathcal{F}$. This verifies Condition B.1(iv).

By definition, $\widetilde{R}_\zeta(h_{F,d})$ in (B.68) does not depend on d for any F. Moreover, for any d and $\widetilde{d}$ with $\|d\| = \infty$ and $\|\widetilde{d}\| = \infty$, by the definition of $R_\zeta(h_{F,d})$ in (B.69),

$$R_\zeta(h_{F,d}) = \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big] = R_\zeta(h_{F,\widetilde{d}}). \tag{B.70}$$

Hence, Condition B.1(vi) is also verified.

We next apply Lemma B.14 to get (B.60) and (B.61) above. By (B.68),

$$R_\zeta(h) - \widetilde{R}_\zeta(h) = \mathbb{E}\big[\min\big\{\overline{\xi}_F' \Upsilon \overline{\xi}_F, \zeta\big\}\big] - \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big] \quad \text{for any } h \in H \tag{B.71}$$

and

$$\begin{aligned} &R_\zeta(h) - \widetilde{R}_\zeta(h) = \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big] - \mathbb{E}\big[\min\big\{\xi_{1,F}' \Upsilon \xi_{1,F}, \zeta\big\}\big] = 0 \\ &\quad \text{for any } h \in H_\infty^*. \end{aligned} \tag{B.72}$$

By Lemma B.14, (B.71), and (B.72), we have

$$\begin{aligned}\mathrm{Asy}\overline{RD}(\widehat{\theta}_{\mathrm{eo}}, \widehat{\theta}_1) &= \lim_{\zeta\to\infty} \max\Big\{\sup_{h\in H}\big[R_\zeta(h) - \widetilde{R}_\zeta(h)\big], \sup_{h\in H_\infty^*}\big[R_\zeta(h) - \widetilde{R}_\zeta(h)\big]\Big\} \\ &= \lim_{\zeta\to\infty} \max\Big\{\sup_{h\in H}\mathbb{E}\big[\min\{\overline{\xi}_F'\Upsilon\overline{\xi}_F, \zeta\} - \min\{\xi_{1,F}'\Upsilon\xi_{1,F}, \zeta\}\big], 0\Big\} \qquad \text{(B.73)}\end{aligned}$$

and

$$\begin{aligned}\mathrm{Asy}\underline{RD}(\widehat{\theta}_{\mathrm{eo}}, \widehat{\theta}_1) &= \lim_{\zeta\to\infty} \min\Big\{\inf_{h\in H}\big[R_\zeta(h) - \widetilde{R}_\zeta(h)\big], \inf_{h\in H_\infty^*}\big[R_\zeta(h) - \widetilde{R}_\zeta(h)\big]\Big\} \\ &= \lim_{\zeta\to\infty} \min\Big\{\inf_{h\in H}\mathbb{E}\big[\min\{\overline{\xi}_F'\Upsilon\overline{\xi}_F, \zeta\} - \min\{\xi_{1,F}'\Upsilon\xi_{1,F}, \zeta\}\big], 0\Big\}, \qquad \text{(B.74)}\end{aligned}$$

which proves (B.60) and (B.61).

In the second step, we show that

$$\lim_{\zeta\to\infty} \max\Big\{\sup_{h\in H}\big[g_\zeta(h)\big], 0\Big\} = \max\Big\{\sup_{h\in H}\big[g(h)\big], 0\Big\}, \quad \text{and} \qquad \text{(B.75)}$$

$$\lim_{\zeta\to\infty} \min\Big\{\inf_{h\in H}\big[g_\zeta(h)\big], 0\Big\} = \min\Big\{\inf_{h\in H}\big[g(h)\big], 0\Big\}. \qquad \text{(B.76)}$$

By Lemma B.16,

$$\lim_{\zeta\to\infty}\sup_{h\in H}\big[g_\zeta(h)\big] = \sup_{h\in H}\big[g(h)\big] \quad \text{and} \quad \lim_{\zeta\to\infty}\inf_{h\in H}\big[g_\zeta(h)\big] = \inf_{h\in H}\big[g(h)\big], \qquad \text{(B.77)}$$

where $\sup_{h\in H}[g(h)]$ and $\inf_{h\in H}[g(h)]$ are finite real numbers. Let $\overline{f}(x) = \max(x, 0)$ and $\underline{f}(x) = \min(x, 0)$. It is clear that $\overline{f}(x)$ and $\underline{f}(x)$ are continuous function on $\mathbb{R}$. The asserted results in (B.75) and (B.76) follow by (B.77), and the continuity of $\overline{f}(x)$ and $\underline{f}(x)$. □

Proof of Theorem 5.2. For any $F \in \mathcal{F}$, define

$$B_F = \big(\Gamma_{2,F} - \Gamma_{1,F}^*\big)'\Upsilon\big(\Gamma_{2,F} - \Gamma_{1,F}^*\big) \quad \text{and} \quad D_F = \big(\Gamma_{2,F} - \Gamma_{1,F}^*\big)'\Upsilon\Gamma_{1,F}^*. \qquad \text{(B.78)}$$

Recall that we have defined $A_F = \Upsilon(\Sigma_{1,F} - \Sigma_{2,F})$ in Theorem 5.2. By the definition of $\overline{\xi}_F$,

$$\mathbb{E}\big[\overline{\xi}_F'\Upsilon\overline{\xi}_F\big] = \mathrm{tr}(\Upsilon\Sigma_{1,F}) + 2\,\mathrm{tr}(A_F)J_{1,F} + \mathrm{tr}(A_F)^2 J_{2,F}, \qquad \text{(B.79)}$$

where

$$\begin{aligned} J_{1,F} &= \mathbb{E}\bigg[\frac{\mathcal{Z}_{d,2,F}' D_F \mathcal{Z}_{d,2,F}}{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \mathrm{tr}(A_F)}\bigg] \quad \text{and} \\ J_{2,F} &= \mathbb{E}\bigg[\frac{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F}}{\big(\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \mathrm{tr}(A_F)\big)^2}\bigg]. \end{aligned} \qquad \text{(B.80)}$$

We provide a upper bound for $J_{1,F}$ defined in (B.80). Define a function

$$\eta(x) \equiv \frac{x}{x' B_F x + \mathrm{tr}(A_F)} \quad \text{for any } x \in \mathbb{R}^{r_2}. \qquad \text{(B.81)}$$

Its derivative is

$$\frac{\partial \eta(x)'}{\partial x} = \frac{1}{x'B_F x + \operatorname{tr}(A_F)} I_{r_2} - \frac{2B_F}{\left(x'B_F x + \operatorname{tr}(A_F)\right)^2} xx'. \tag{B.82}$$

Then $J_{1,F} = \mathbb{E}[\eta(\mathcal{Z}_{d,2,F})' D_F \mathcal{Z}_{d,2,F}]$. Note that $D_F \mathcal{Z}_{d,2,F} = D_F \mathcal{Z}_{2,F}$ by construction because the last r^* columns of $\Gamma^*_{1,F}$ are zeros. Applying Lemma B.9 yields

$$\begin{aligned} \operatorname{tr}(D_F \Omega_{2,F}) &= \operatorname{tr}\left(\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right)' \Upsilon \Gamma^*_{1,F} \Omega_{2,F}\right) \\ &= \operatorname{tr}\left(\Upsilon\left(\Gamma^*_{1,F} \Omega_{2,F} \Gamma'_{2,F} - \Gamma^*_{1,F} \Omega_{2,F} \Gamma^*_{1,F}\right)\right) \\ &= \operatorname{tr}\left(\Upsilon(\Sigma_{2,F} - \Sigma_{1,F})\right) = -\operatorname{tr}(A_F). \end{aligned} \tag{B.83}$$

By Lemma 1 of Hansen (2016), which is a matrix version of the Stein's lemma (Stein (1981)),

$$J_{1,F} = \mathbb{E}\left(\eta(\mathcal{Z}_{d,2,F})' D_F \mathcal{Z}_{d,2,F}\right) = \mathbb{E}\left[\operatorname{tr}\left(\frac{\partial \eta(\mathcal{Z}_{d,2,F})'}{\partial x} D_F \Omega_{2,F}\right)\right]. \tag{B.84}$$

Plugging (B.81)–(B.83) into (B.84), we have

$$\begin{aligned} J_{1,F} &= \mathbb{E}\left[\frac{\operatorname{tr}(D_F \Omega_{2,F})}{\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] - 2\mathbb{E}\left[\frac{\operatorname{tr}\left(B_F \mathcal{Z}_{d,2,F} \mathcal{Z}'_{d,2,F} D_F \Omega_{2,F}\right)}{\left(\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right)^2}\right] \\ &= \mathbb{E}\left[\frac{-\operatorname{tr}(A_F)}{\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] + 2\mathbb{E}\left[\frac{-\mathcal{Z}'_{d,2,F} D_F \Omega_{2,F} B_F \mathcal{Z}_{d,2,F}}{\left(\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right)^2}\right], \end{aligned} \tag{B.85}$$

where the second equality is by (B.83). By definition and Lemma B.9

$$\begin{aligned} &-\mathcal{Z}'_{d,2,F} D_F \Omega_{2,F} B_F \mathcal{Z}_{d,2,F} \\ &\quad = -\mathcal{Z}'_{d,2,F}\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right)' \Upsilon \Gamma^*_{1,F} \Omega_{2,F}\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right)' \Upsilon\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right) \mathcal{Z}_{d,2,F} \\ &\quad = \mathcal{Z}'_{d,2,F}\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right)' \Upsilon(\Sigma_{1,F} - \Sigma_{2,F}) \Upsilon\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right) \mathcal{Z}_{d,2,F} \\ &\quad \leq \rho_{\max}\left(\Upsilon^{1/2}(\Sigma_{1,F} - \Sigma_{2,F}) \Upsilon^{1/2}\right)\left(\mathcal{Z}'_{d,2,F}\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right)' \Upsilon\left(\Gamma_{2,F} - \Gamma^*_{1,F}\right) \mathcal{Z}_{d,2,F}\right) \\ &\quad = \rho_{\max}(A_F) \mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F}, \end{aligned} \tag{B.86}$$

where the last equality is by $\rho_{\max}(\Upsilon^{1/2}(\Sigma_{1,F} - \Sigma_{2,F})\Upsilon^{1/2}) = \rho_{\max}(\Upsilon(\Sigma_{1,F} - \Sigma_{2,F}))$. Combining the results in (B.85) and (B.86), we get

$$\begin{aligned} J_{1,F} &\leq \mathbb{E}\left[\frac{-\operatorname{tr}(A_F)}{\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] + 2\mathbb{E}\left[\frac{\rho_{\max}(A_F) \mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F}}{\left(\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right)^2}\right] \\ &= \mathbb{E}\left[\frac{-\operatorname{tr}(A_F)}{\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] \\ &\quad + 2\mathbb{E}\left[\frac{\left[\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A)\right] \rho_{\max}(A_F) - \operatorname{tr}(A_F) \rho_{\max}(A_F)}{\left(\mathcal{Z}'_{d,2,F} B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right)^2}\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{2\rho_{\max}(A_F) - \operatorname{tr}(A_F)}{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] \\
&\quad - \mathbb{E}\left[\frac{2\rho_{\max}(A_F)\operatorname{tr}(A_F)}{\left(\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right)^2}\right]. \qquad \text{(B.87)}
\end{aligned}
$$

Next, note that

$$
\begin{aligned}
J_{2,F} &= \mathbb{E}\left[\frac{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F}}{\left|\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right|^2}\right] \\
&= \mathbb{E}\left[\frac{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F) - \operatorname{tr}(A_F)}{\left|\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right|^2}\right] \\
&= \mathbb{E}\left[\frac{1}{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] \\
&\quad - \mathbb{E}\left[\frac{\operatorname{tr}(A_F)}{\left|\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right|^2}\right]. \qquad \text{(B.88)}
\end{aligned}
$$

Combining (B.79), (B.87), (B.88), and the definition of $g(h)$ (in Theorem 5.1), we obtain that

$$
\begin{aligned}
g(h_{d,F}) &= 2\operatorname{tr}(A_F)J_{1,F} + \operatorname{tr}(A_F)^2 J_{2,F} \\
&\leq 2\operatorname{tr}(A_F)\left(\mathbb{E}\left[\frac{2\rho_{\max}(A_F) - \operatorname{tr}(A_F)}{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] - \mathbb{E}\left[\frac{2\operatorname{tr}(A_F)\rho_{\max}(A_F)}{\left|\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right|^2}\right]\right) \\
&\quad + \operatorname{tr}(A)^2\left(\mathbb{E}\left[\frac{1}{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] - \mathbb{E}\left[\frac{\operatorname{tr}(A_F)}{\left|\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right|^2}\right]\right) \\
&= \mathbb{E}\left[\frac{\operatorname{tr}(A_F)\left(4\rho_{\max}(A_F) - \operatorname{tr}(A_F)\right)}{\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)}\right] \\
&\quad - \mathbb{E}\left[\frac{\operatorname{tr}(A_F)^2\left(4\rho_{\max}(A_F) + \operatorname{tr}(A_F)\right)}{\left|\mathcal{Z}_{d,2,F}' B_F \mathcal{Z}_{d,2,F} + \operatorname{tr}(A_F)\right|^2}\right]. \qquad \text{(B.89)}
\end{aligned}
$$

For all G_2 and Ω_2 such that $h = (d, \operatorname{vec}(G_2)', \operatorname{vech}(\Omega_2)') \in H$, we have $G_2 = G_{2,F}$ and $\Omega_2 = \Omega_{2,F}$ for some $F \in \mathcal{F}$ by the definition of H. If $\operatorname{tr}(A_F) > 0$, then $\rho_{\max}(A_F) > 0$, and thus the second term in the right-hand side of the last equality of (B.89) will be negative. If in addition $\operatorname{tr}(A_F) \geq 4\rho_{\max}(A_F)$, then the first term in the right-hand side of the last equality of (B.89) will be nonnegative. As a result, when $\operatorname{tr}(A_F) > 0$ and $4\rho_{\max}(A_F) - \operatorname{tr}(A_F) \leq 0$ for $\forall F \in \mathcal{F}$, we have $\sup_{h\in H}[g(h)] < 0$. This combined with Theorem 5.1 implies the results of this theorem. □

Appendix C: Supplementary simulation results

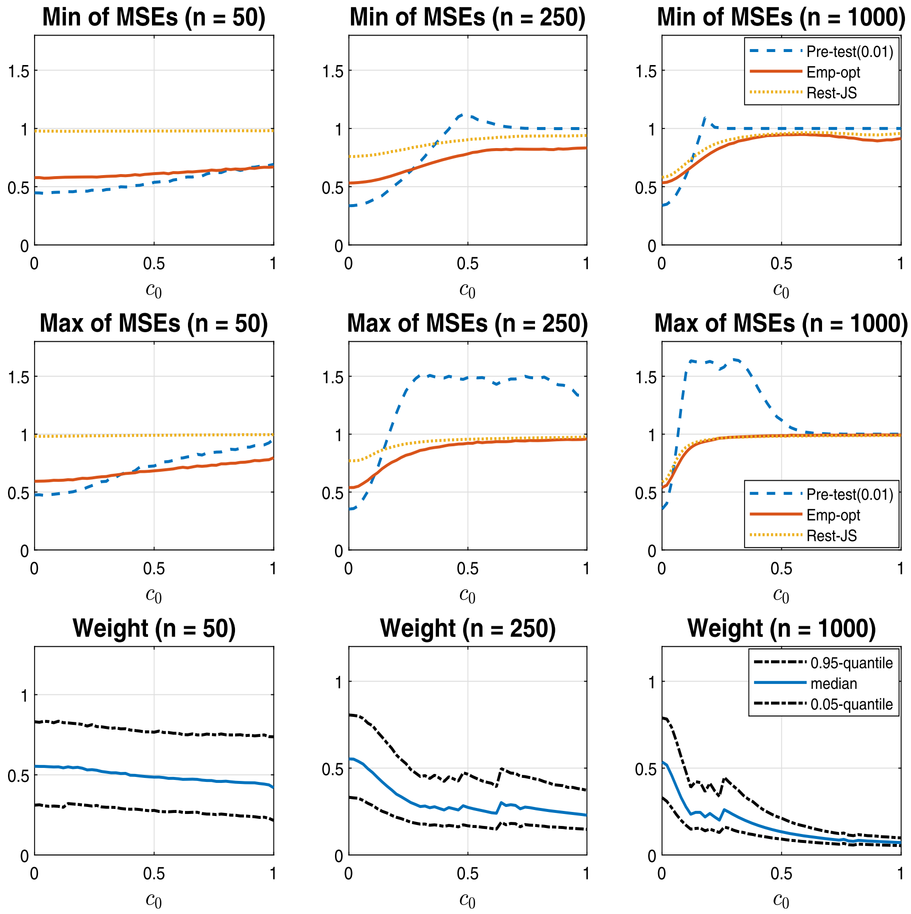

Figure C.1. Finite sample MSEs of the pre-test and averaging GMM estimators in S1. *Note*: "Pre-test(0.01)" refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; "Emp-opt" refers to the averaging GMM estimator based on the empirical optimal weight; "Rest-JS" refers to the averaging estimators based on the restricted James–Stein weight, respectively.

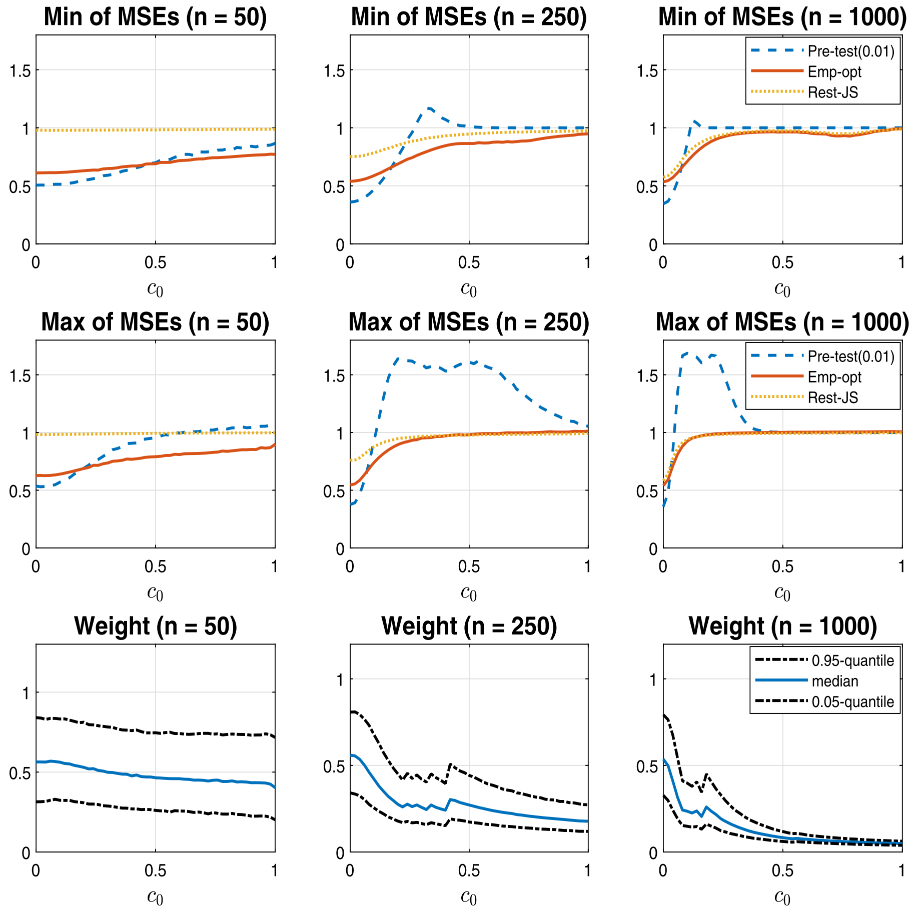

Figure C.2. Finite sample MSEs of the pre-test and averaging GMM estimators in S2. *Note*: "Pre-test(0.01)" refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; "Emp-opt" refers to the averaging GMM estimator based on the empirical optimal weight; "Rest-JS" refers to the averaging estimators based on the restricted James–Stein weight, respectively.

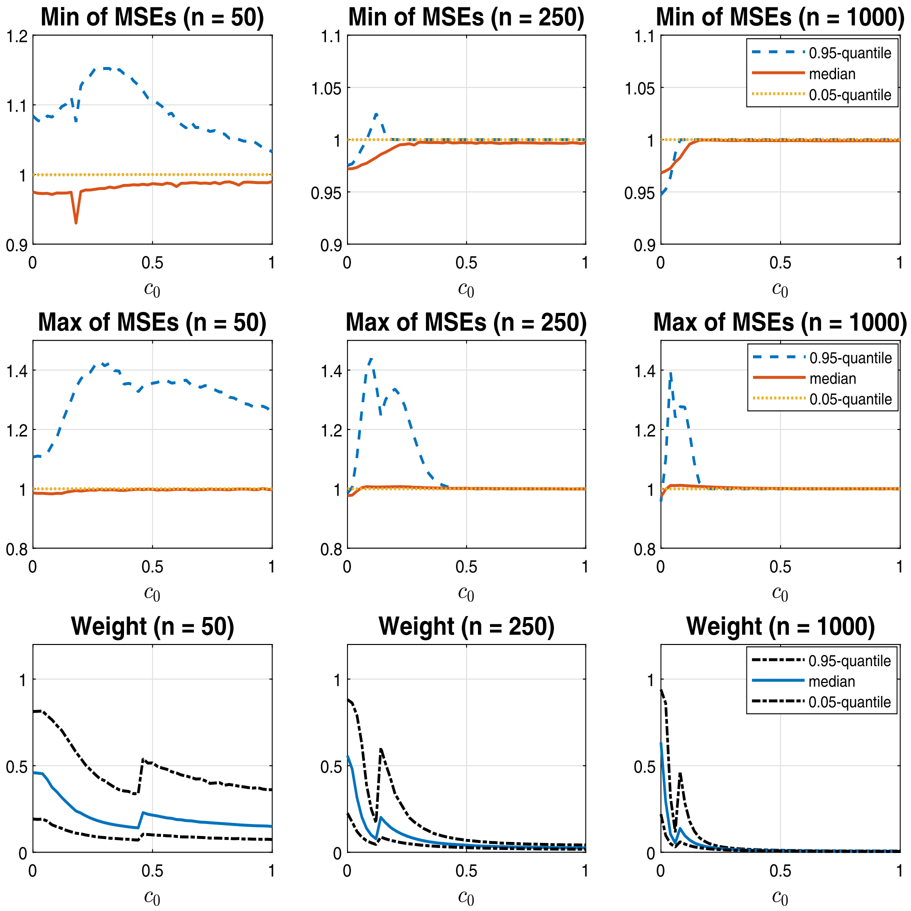

FIGURE C.3. Finite sample MSEs of the pre-test and averaging GMM estimators in S3. *Note*: "Pre-test(0.01)" refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; "Emp-opt" refers to the averaging GMM estimator based on the empirical optimal weight; "Rest-JS" refers to the averaging estimators based on the restricted James–Stein weight, respectively.

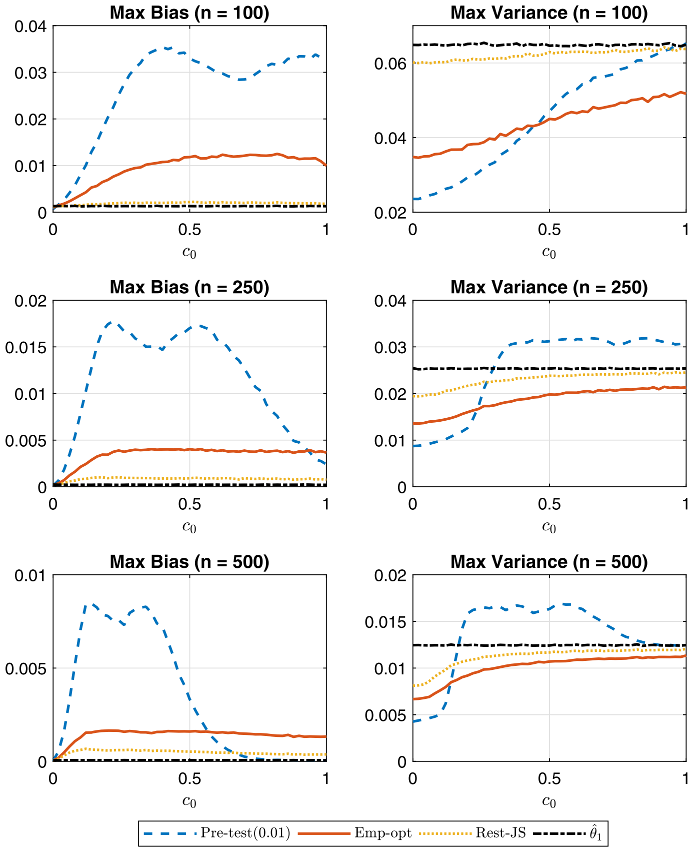

FIGURE C.4. Finite sample biases and variances in S1. *Note*: “Pre-test(0.01)” refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; “Emp-opt” refers to the averaging GMM estimator based on the empirical optimal weight; “Rest-JS” refers to the averaging estimators based on the restricted James–Stein weight, respectively.

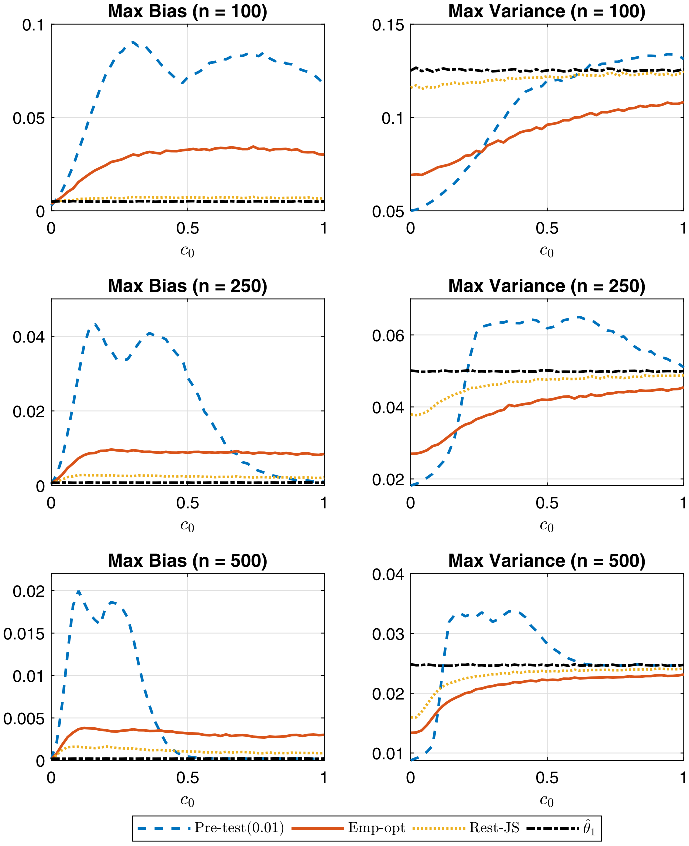

Figure C.5. Finite sample biases and variances in S2. *Note*: "Pre-test(0.01)" refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; "Emp-opt" refers to the averaging GMM estimator based on the empirical optimal weight; "Rest-JS" refers to the averaging estimators based on the restricted James–Stein weight, respectively.

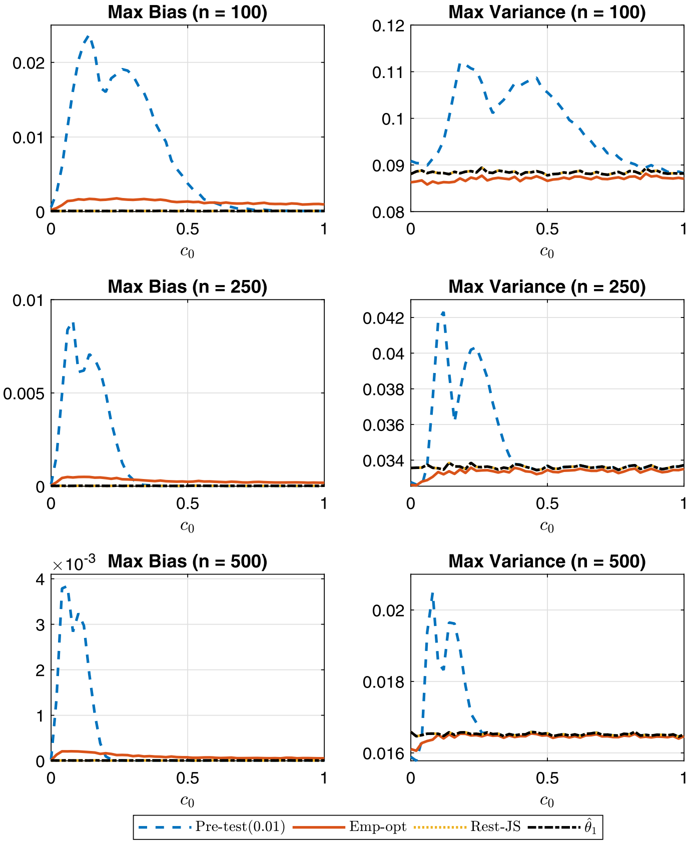

FIGURE C.6. Finite sample biases and variances in S3. "Pre-test(0.01)" refers to the pre-test GMM estimator based on the J-test with nominal size 0.01; "Emp-opt" refers to the averaging GMM estimator based on the empirical optimal weight; "Rest-JS" refers to the averaging estimators based on the restricted James–Stein weight, respectively.

References

Andrews, D. W. K. (1999), "Consistent moment selection procedures for generalized method of moments estimation." *Econometrica*, 67 (3), 543–563. [937]

Andrews, D. W. K. and X. Cheng (2013), "Maximum likelihood estimation and uniform inference with sporadic identification failure." *Journal of Econometrics*, 173 (1), 36–56. [956]

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2011), "Generic results for establishing the asymptotic size of confidence sets and tests." Cowles Foundation Discussion Paper, No. 1813. [936, 960]

Andrews, D. W. K. and B. Lu (2001), "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models." *Journal of Econometrics*, 101 (1), 123–164. [937]

Ashley, R. (2009), "Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis." *Journal of Applied Econometrics*, 24, 325–337. [937]

Berkowitz, D., M. Caner, and Y. Fang (2012), "The validity of instruments revisited." *Journal of Econometrics*, 166 (2), 255–266. [937]

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997), "Model selection: An integral part of inference." *Biometrics*, 53, 603–618. [936]

Charkhi, A., G. Claeskens, and B. Hansen (2016), "Minimum mean squared error model averaging in likelihood models." *Statistica Sinica*, 26, 809–840. [943]

Cheng, X. and B. Hansen (2015), "Forecasting with factor-augmented regression: A frequentist model averaging approach." *Journal of Econometrics*, 186 (2), 280–293. [936]

Cheng, X. and Z. Liao (2015), "Select the valid and relevant moments: An information-based LASSO for GMM with many moments." *Journal of Econometrics*, 186 (2), 443–464. [937]

Cheng, X., Z. Liao, and R. Shi (2019), "Supplement to 'On uniform asymptotic risk of averaging GMM estimators'." *Quantitative Economics Supplemental Material*, 10, https://doi.org/10.3982/QE711. [934, 955]

Claeskens, G. and R. J. Carroll (2007), "An asymptotic theory for model selection inference in general semiparametric problems." *Biometrika*, 94, 249–265. [936]

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012), "Plausibly exogenous." *Review of Economics and Statistics*, 94 (1), 260–272. [937]

DiTraglia, F. (2016), "Using invalid instruments on purpose: Focused moment selection and averaging for GMM." *Journal of Econometrics*, 195 (2), 187–208. [936]

Doko Tchatoka, F. and J.-M. Dufour (2008), "Instrument endogeneity and identification-robust tests: Some analytical results." *Journal of Statistical Planning and Inference*, 138 (9), 2649–2661. [937]

Doko Tchatoka, F. and J.-M. Dufour (2014), "Identification-robust inference for endogeneity parameters in linear structural models." *Econometrics Journal*, 17, 165–187. [937]

Eichenbaum, M., L. Hansen, and K. Singleton (1988), "A time series analysis of representative agent models of consumption and leisure choice under uncertainty." *Quarterly Journal of Economics*, 103 (1), 51–78. [937]

Green, E. J. and W. E. Strawderman (1991), "A James-Stein type estimator for combining unbiased and possibly biased estimators." *Journal of the American Statistical Association*, 86 (416), 1001–1006. [935, 954, 955]

Guggenberger, P. (2012), "On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption." *Econometric Theory*, 28 (2), 387–421. [937]

Hansen, B. E. (2007), "Least squares model averaging." *Econometrica*, 75, 1175–1189. [936]

Hansen, B. E. (2016), "Efficient shrinkage in parametric models." *Journal of Econometrics*, 190 (1), 115–132. [935, 946, 968]

Hansen, B. E. (2017), "A Stein-like 2SLS estimator." *Econometric Reviews*, 36, 840–852. [936]

Hansen, B. E. and J. Racine (2012), "Jackknife model averaging." *Journal of Econometrics*, 167, 38–46. [936]

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators." *Econometrica*, 50, 1029–1054. [932, 933, 937]

Hjort, N. L. and G. Claeskens (2003), "Frequentist model average estimators." *Journal of the American Statistical Association*, 98, 879–899. [936]

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, 1449–1464. [936]

Hong, H., B. Preston, and M. Shum (2003), "Generalized empirical likelihood-based model selection criteria for moment condition models." *Econometric Theory*, 19 (6), 923–943. [937]

James, W. and C. Stein (1961), "Estimation with quadratic loss." *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, 1, 361–380. [935]

Judge, G. C. and R. C. Mittelhammer (2004), "A semiparametric basis for combining estimating problems under quadratic loss." *Journal of American Statistical Association*, 99, 479–487. [935, 954, 955]

Judge, G. C. and R. C. Mittelhammer (2007), "Estimation and inference in the case of competing sets of estimating equations." *Journal of Econometrics*, 138, 513–531. [935]

Kabaila, P. (1995), "The effect of model selection on confidence regions and prediction regions." *Econometric Theory*, 11, 537–549. [933]

Kabaila, P. (1998), "Valid confidence intervals in regression after variable selection." *Econometric Theory*, 14, 463–482. [933]

Kang, H., A. Zhang, T. T. Cai, and D. S. Small (2016), "Instrumental variables estimation with some invalid instruments and its application to mendelian randomization." *Journal of the American Statistical Association*, 111 (513), 132–144. [937]

Kim, T. H. and H. White (2001), "James–Stein type estimators in large samples with application to the least absolute deviation estimator." *Journal of the American Statistical Association*, 96, 697–705. [935, 954, 955]

Kolesar, M., R. Chetty, J. Friedman, E. Glaeser, and G. Imbens (2015), "Identification and inference with many invalid instruments." *Journal of Business Economics and Statistics*, 33 (4), 474–484. [937]

Leeb, H. and B. M. Pötscher (2005), "Model selection and inference: Facts and fiction." *Econometric Theory*, 21, 21–59. [933]

Leeb, H. and B. M. Pötscher (2008), "Sparse estimators and the oracle property, or the return of the Hodge's estimator." *Journal of Econometrics*, 142 (1), 201–211. [933, 936]

Lehmann, E. L. and G. Casella (1998), *Theory of Point Estimation*. Springer-Verlag, New York, NY. [936]

Leung, G. and A. Barron (2006), "Information theory and mixing least-square regressions." *IEEE Transactions on Information Theory*, 52, 3396–3410. [937]

Liao, Z. (2013), "Adaptive GMM shrinkage estimation with consistent moment selection." *Econometric Theory*, 29, 1–48. [937]

Liu, C.-A. (2015), "Distribution theory of the least squares averaging estimator." *Journal of Econometrics*, 186 (1), 142–159. [943]

Lu, X. and L. Su (2015), "Jackknife model averaging for quantile regressions." *Journal of Econometrics*, 188, (1), 40–58. [936]

Mittelhammer, R. C. and G. C. Judge (2005), "Combining estimators to improve structural model estimation and inference under quadratic loss." *Journal of Econometrics*, 128, (1), 1–29. [935, 954, 955]

Moon, H. R. and F. Schorfheide (2009), "Estimation with overidentifying inequality moment conditions." *Journal of Econometrics*, 153 (2), 136–154. [937]

Nevo, A. and A. Rosen (2012), "Identification with imperfect instruments." *Review of Economics and Statistics*, 93 (3), 659–671. [937]

Pötscher, B. M. (1991), "Effects of model selection on inference." *Econometric Theory*, 7, 163–185. [933]

Pötscher, B. M. (2006), "The distribution of model averaging estimators and an impossibility result regarding its estimation." In *Time Series and Related Topics: in Memory of Ching-Zong Wei* (H.-C. Ho, C.-K. Ing, and T.-L. Lai, eds.), IMS Lecture Notes and Monograph Series, Vol. 52, 113–129. [954]

Sargan, J. (1958), "The estimation of economic relationships using instrumental variables." *Econometrica*, 26 (3), 393–415. [937]

Shibata, R. (1986), "Consistency of model selection and parameter estimation." *Journal of Applied Probability, special*, 23A, 127–141. [933]

Small, D. S. (2007), "Sensitivity analysis for instrumental variables regression with overidentifying restrictions." *Journal of the American Statistical Association*, 102 (479), 1049–1058. [937]

Stein, C. M. (1956), "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1, 197–206. [935]

Stein, C. M. (1981), "Estimation of the mean of a multivariate normal distribution." *The Annals of Statistics*, 1135–1151. [968]

Van der Vaart, A. W. (1998), *Asymptotic Statistics*. Cambridge University Press. [966]

Yang, Y. (2000), "Combining different regression procedures for adaptive regression." *Journal of Multivariate Analysis*, 74, 135–161. [937]

Yang, Y. (2003), "Regression with multiple candidate models: Selecting or mixing?" *Statistica Sinica*, 13, 783–809. [937]

Yang, Y. (2004), "Combining forecasting procedures: Some theoretical results." *Econometric Theory*, 20, 176–222. [937]

Yang, Y. (2005), "Can the strengths of AIC and BIC be shared?—A conflict between model identification and regression estimation." *Biometrika*, 92, 937–950. [936]

Co-editor Karl Schmedders handled this manuscript.

Manuscript received 3 May, 2016; final version accepted 27 November, 2018; available online 9 January, 2019.