

Penn Institute for Economic Research  
Department of Economics  
University of Pennsylvania  
3718 Locust Walk  
Philadelphia, PA 19104-6297  
[pier@econ.upenn.edu](mailto:pier@econ.upenn.edu)  
<http://economics.sas.upenn.edu/pier>

## *PIER Working Paper 14-037*

“Using Invalid Instruments on Purpose: Focused  
Moment Selection and Averaging for GMM”

by

Francis J. DiTraglia

<http://ssrn.com/abstract=2516255>

# Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM\*

Francis J. DiTraglia  
University of Pennsylvania

First Version: November 9, 2011

Current Version: August 4, 2014

## Abstract

In finite samples, the use of a slightly endogenous but highly relevant instrument can reduce mean-squared error (MSE). Building on this observation, I propose a moment selection criterion for GMM in which moment conditions are chosen based on the MSE of their associated estimators rather than their validity: the focused moment selection criterion (FMSC). I then show how the framework used to derive the FMSC can address the problem of inference post-moment selection. Treating post-selection estimators as a special case of moment-averaging, in which estimators based on different moment sets are given data-dependent weights, I propose a simulation-based procedure to construct valid confidence intervals for a variety of formal and informal moment-selection procedures. Both the FMSC and confidence interval procedure perform well in simulations. I conclude with an empirical example examining the effect of instrument selection on the estimated relationship between malaria transmission and per-capita income.

## 1 Introduction

In finite samples, the addition of a slightly endogenous but highly relevant instrument can yield a substantial improvement, reducing estimator variance by far more than bias is increased. Building on this observation, I propose a novel moment selection criterion for generalized method of moments (GMM) estimation: the focused moment

---

\*I thank Aislinn Bohren, Xu Cheng, Gerda Claeskens, Bruce Hansen, Byunghoon Kang, Toru Kitagawa, Hannes Leeb, Serena Ng, Alexei Onatski, Hashem Pesaran, Benedikt Pötscher, Frank Schorfheide, Neil Shephard, Richard J. Smith, Stephen Thiele, Melvyn Weeks, as well as seminar participants at Cambridge, the University of Vienna, Queen Mary, St Andrews, George Washington, UPenn, Columbia, Oxford, and the 2011 Econometric Society European Meetings for their many helpful comments and suggestions. I thank Kai Carstensen for providing data for my empirical example.

selection criterion (FMSC). Rather than selecting only valid moment conditions, the FMSC chooses from a set of potentially mis-specified moment conditions to yield the minimum asymptotic mean squared error (AMSE) GMM estimator of a user-specified target parameter. To ensure a meaningful bias-variance tradeoff in the limit, I employ a drifting asymptotic framework in which mis-specification, while present for any fixed sample size, vanishes asymptotically. In the presence of such *locally mis-specified* moment conditions, GMM remains consistent although, centered and rescaled, its limiting distribution displays an asymptotic bias. Adding an additional mis-specified moment condition introduces a further source of bias while reducing asymptotic variance. The idea behind the FMSC is to trade off these two effects in the limit as an approximation to finite sample behavior.<sup>1</sup>

I consider a setting in which two blocks of moment conditions are available: one that is assumed correctly specified, and another that may not be. This is intended to mimic the situation faced by an applied researcher who begins with a “baseline” set of maintained assumptions and must decide whether to impose any of a collection of stronger but also more controversial “suspect” assumptions. To be credible, the baseline assumptions cannot be particularly strong. As such, imposing one, or more, of the stronger suspect assumptions could yield more precise estimates, provided that the added assumption is at least *nearly* correct. When the (correctly specified) baseline moment conditions identify the model, the FMSC provides an asymptotically unbiased estimator of AMSE, allowing us to carry out risk-based selection over the suspect moment conditions. When this is not the case, it remains possible to use the AMSE framework to carry out a sensitivity analysis.

Continuing under the local mis-specification assumption, I show how the framework employed in the derivation of the FMSC can be used to address the important problem of inference post-moment selection. I treat post-selection estimators as a special case of moment averaging: combining estimators based on different moment sets with data-dependent weights. By deriving the limiting distribution of this class of estimators in general, I propose a simulation-based procedure for constructing valid confidence intervals. This technique can be applied to a variety of formal moment averaging and post-selection estimators, such as the FMSC, as well as pre-test estimators based on the  $J$ - and Durbin-Hausman-Wu-statistic (DHW).

While the methods described here apply to general GMM models, subject to standard regularity conditions, I focus on two simple but empirically relevant examples: choosing between ordinary least squares (OLS) and two-stage least squares (TSLS) estimators, and selecting instruments in linear instrumental variables (IV) models. In the OLS versus TSLS example the FMSC takes a particularly intuitive and transparent form, providing a risk-based justification for the DHW test, and leading to a novel “minimum-AMSE” estimator that optimally combines the information contained in

---

<sup>1</sup>When finite-sample mean-squared error (MSE) is undefined or infinite, AMSE comparisons remain meaningful. In this case, one can view AMSE as the limit of a sequence of “trimmed” squared error loss functions, as in Hansen (2013). Trimmed MSE is always well-defined and the trimming fraction can be made asymptotically negligible.

the OLS and TSLS estimators. In simulation studies for both examples, the FMSC and confidence interval procedures perform well. The cost, in terms of estimator risk, of using the FMSC rather than simply carrying out estimation under the baseline moment conditions is small; the gains, on the other hand, can be substantial. I conclude with an empirical application from development economics, exploring the effect of instrument selection on the estimated relationship between malaria transmission and per capita income.

My approach to moment selection under mis-specification is inspired by the focused information criterion of [Claeskens and Hjort \(2003\)](#), a model selection criterion for models estimated by maximum likelihood. Like them, I allow for mis-specification and study AMSE-based selection in a drifting asymptotic framework. In contradistinction, however, I consider moment rather than model selection, and general GMM rather than maximum likelihood estimators. Although developed independently of the FIC, [Schorfheide \(2005\)](#) uses a similar approach to select over forecasts constructed from mis-specified vector autoregression models. Mine is by no means the first paper to consider GMM asymptotics under locally mis-specified moment conditions, an idea that dates at least as far back as [Newey \(1985\)](#). The idea of using this framework for AMSE-based moment selection, however, is new to the literature.

The FMSC differs substantially from existing criteria for moment selection under mis-specification whose primary aim is to select all correctly specified moment conditions while eliminating all invalid ones with probability approaching one in the limit. This idea begins with [Andrews \(1999\)](#), who proposes a family of consistent moment selection criteria for GMM by adding an appropriate penalty term to the  $J$ -statistic. [Andrews and Lu \(2001\)](#) extend this work to allow simultaneous GMM moment and model selection, while [Hong et al. \(2003\)](#) derive analogous results for generalized empirical likelihood. More recently, [Liao \(2013\)](#) proposes a shrinkage procedure for simultaneous GMM moment selection and estimation. Given a set of correctly specified moment conditions that identifies the model, this method consistently chooses all valid conditions from a second set of potentially mis-specified conditions. In contrast to these proposals, which examine only the validity of the moment conditions under consideration, the FMSC balances validity against relevance to minimize AMSE. Although [Hall and Peixe \(2003\)](#) and [Cheng and Liao \(2013\)](#) do consider relevance, their aim is to avoid including redundant moment conditions after consistently eliminating invalid ones. In contrast to the FMSC, they do not allow for the intentional inclusion of a slightly invalid but highly relevant instrument to reduce AMSE. Some other papers that propose choosing, or combining, instruments to minimize MSE include [Donald and Newey \(2001\)](#), [Donald et al. \(2009\)](#), and [Kuersteiner and Okui \(2010\)](#). Unlike the FMSC, however, these proposals consider the *higher-order* bias that arises from including many valid instruments rather than the first-order bias that arises from the use of invalid instruments.

Another important difference between the FMSC and the other proposals from the literature is the “F” – focus. Indeed, rather than a single moment selection criterion, the FMSC is really a method of constructing application-specific moment selection criteria.

Consider, for example, a simple dynamic panel model. If your target parameter is a long-run effect while mine is a contemporaneous effect, there is no reason to suppose *a priori* that we should use the same moment conditions in estimation, even if we share the same model and dataset. The FMSC explicitly takes this difference of research goals into account, unlike other moment selection criteria from the literature.

Like Akaike’s Information Criterion (AIC), the FMSC is a conservative rather than consistent selection procedure. As such, unlike the consistent moment selection criteria described above, it is random *even in the limit*. This may sound disconcerting. Although consistency is a crucial minimal property in many settings, the situation is more complex for model and moment selection. The problem is that consistent and conservative selection procedures have different strengths, but these strengths cannot be combined [Yang \(2005\)](#). Which should be preferred depends on the application one has in mind. If our goal is to evaluate the validity of an assumption from economic theory, consistent selection is the natural choice, as it will choose only correctly specified moment conditions with probability approaching one in the limit. The motivation behind the FMSC, however, is to minimize estimator risk. With this aim in mind, consistent selection criteria suffer from a serious defect: in general, unlike conservative criteria, they exhibit *unbounded* minimax risk ([Leeb and Pötscher, 2008](#)). Moreover, as discussed in more detail below, the asymptotics of consistent selection paint a seriously misleading picture of the effects of moment selection on inference. For these reasons, the fact that the FMSC is conservative rather than consistent is a “feature” rather than a “bug.”

Because it studies inference post-moment selection, this paper relates to a vast literature on “pre-test” and post-selection inference. [Leeb and Pötscher \(2005, 2009\)](#) give a theoretical overview of the consequences of model selection for subsequent inference, while [Demetrescu et al. \(2011\)](#) illustrate them in a simulation study. There are several proposals to construct valid confidence intervals post-model selection, including [Kabaila \(1998\)](#), [Hjort and Claeskens \(2003\)](#) and [Kabaila and Leeb \(2006\)](#). To my knowledge, however, this is the first paper to treat the problem in general for post-moment selection and moment average estimators in the presence of mis-specification.<sup>2</sup> While I developed the simulation-based, two-stage confidence interval procedure described below by analogy to [Claeskens and Hjort \(2008b\)](#), [Leeb and Pötscher \(2014\)](#) kindly pointed out that similar constructions have appear in [Loh \(1985\)](#), [Berger and Boos \(1994\)](#), and [Silvapulle \(1996\)](#). More recently, [McCloskey \(2012\)](#) takes a similar approach to study a class of non-standard testing problems.

The idea of treating post-moment selection estimators as a special case example of moment averaging, is adapted from the frequentist model average estimators of [Hjort and Claeskens \(2003\)](#). Two other papers that consider weighting estimators based on different moment conditions are [Xiao \(2010\)](#) and [Chen et al. \(2009\)](#). Whereas these papers combine estimators based on valid moment conditions to achieve a minimum

---

<sup>2</sup>Some related results appear in [Berkowitz et al. \(2012\)](#), [Guggenberger \(2010\)](#), [Guggenberger \(2012\)](#), and [Guggenberger and Kumar \(2012\)](#).

variance estimator, however, I combine estimators based on potentially invalid conditions to minimize AMSE. A similar idea underlies the combined moments (CM) estimator of [Judge and Mittelhammer \(2007\)](#), who emphasize that incorporating the information from an incorrect specification could lead to favorable bias-variance trade-off. Their proposal uses a Cressie-Read divergence measure to combine the information from competing moment specifications, for example OLS versus TSLS, yielding a data-driven compromise estimator. Unlike the FMSC, however, the CM estimator is not targeted to a particular research goal and does not explicitly aim to minimize AMSE.

An important limitation of the results presented here is that they are based upon the assumption of strong identification and a fixed number of moment conditions. I do not, for example, consider weak- or many-instrument asymptotics. Extending the idea behind the FMSC to a more general framework that allows for weak identification and a growing number of moment conditions is a challenging and important topic that I leave for future research.

The remainder of the paper is organized as follows. Section 2 describes the local mis-specification framework, introduces the two running examples, OLS versus TSLS and choosing instruments, and gives the main limiting results used later in the paper. Section 3 derives FMSC as an asymptotically unbiased estimator of AMSE, presents specialized results for the two running examples, and examines their performance in a Monte Carlo experiment. Section 4 describes a simulation-based procedure to construct valid confidence intervals for moment average estimators and examines its performance in a Monte Carlo experiment. Section 5 presents the empirical application and Section 6 concludes. Proofs and computational details appear in the Appendix.

## 2 Assumptions and Asymptotic Framework

### 2.1 Local Mis-Specification

Let  $f(\cdot, \cdot)$  be a  $(p + q)$ -vector of moment functions of a random vector  $Z$  and an  $r$ -dimensional parameter vector  $\theta$ , partitioned according to  $f(\cdot, \cdot) = (g(\cdot, \cdot)', h(\cdot, \cdot)')$  where  $g(\cdot, \cdot)$  and  $h(\cdot, \cdot)$  are  $p$ - and  $q$ -vectors of moment functions. The moment condition associated with  $g(\cdot, \cdot)$  is assumed to be correct whereas that associated with  $h(\cdot, \cdot)$  is locally mis-specified. More precisely,

**Assumption 2.1** (Local Mis-Specification). *Let  $\{Z_{ni}: 1 \leq i \leq n, n = 1, 2, \dots\}$  be a triangular array of random vectors defined on a probability space  $(\Upsilon, \mathcal{F}, \mathbb{P})$  satisfying*

- (a)  $E[g(Z_{ni}, \theta_0)] = 0$ ,
- (b)  $E[h(Z_{ni}, \theta_0)] = n^{-1/2}\tau$ , where  $\tau$  is an unknown constant vector,
- (c)  $\{f(Z_{ni}, \theta_0): 1 \leq i \leq n, n = 1, 2, \dots\}$  is uniformly integrable, and
- (d)  $Z_{ni} \rightarrow_d Z_i$ , where the  $Z_i$  are identically distributed.

For any fixed sample size  $n$ , the expectation of  $h$  evaluated at the true parameter value  $\theta_0$  depends on the unknown constant vector  $\tau$ . Unless all components of  $\tau$  are zero, some of the moment conditions contained in  $h$  are mis-specified. In the limit however, this mis-specification vanishes, as  $\tau/\sqrt{n}$  converges to zero. Uniform integrability combined with weak convergence implies convergence of expectations, so that  $E[g(Z_i, \theta_0)] = 0$  and  $E[h(Z_i, \theta_0)] = 0$ . Because the limiting random vectors  $Z_i$  are identically distributed, I suppress the  $i$  subscript and simply write  $Z$  to denote their common marginal law, e.g.  $E[h(Z, \theta_0)] = 0$ . It is important to note that local mis-specification is *not* intended as a literal description of real-world datasets: it is merely a device that gives asymptotic bias-variance trade-off that mimics the finite-sample intuition.

## 2.2 Candidate GMM Estimators

Define the sample analogue of the expectations in Assumption 2.1 as follows:

$$f_n(\theta) = \frac{1}{n} \sum_{i=1}^n f(Z_{ni}, \theta) = \begin{bmatrix} g_n(\theta) \\ h_n(\theta) \end{bmatrix} = \begin{bmatrix} n^{-1} \sum_{i=1}^n g(Z_{ni}, \theta) \\ n^{-1} \sum_{i=1}^n h(Z_{ni}, \theta) \end{bmatrix}$$

where  $g_n$  is the sample analogue of the correctly specified moment conditions and  $h_n$  is that of the (potentially) mis-specified moment conditions. A candidate GMM estimator  $\widehat{\theta}_S$  uses some subset  $S$  of the moment conditions contained in  $f$  in estimation. Let  $|S|$  denote the number of moment conditions used and suppose that  $|S| > r$  so the GMM estimator is unique.<sup>3</sup> Let  $\Xi_S$  be the  $|S| \times (p+q)$  *moment selection matrix* corresponding to  $S$ . That is,  $\Xi_S$  is a matrix of ones and zeros arranged such that  $\Xi_S f_n(\theta)$  contains only the sample moment conditions used to estimate  $\widehat{\theta}_S$ . Thus, the GMM estimator of  $\theta$  based on moment set  $S$  is given by

$$\widehat{\theta}_S = \arg \min_{\theta \in \Theta} [\Xi_S f_n(\theta)]' \widetilde{W}_S [\Xi_S f_n(\theta)].$$

where  $\widetilde{W}_S$  is an  $|S| \times |S|$ , positive semi-definite weight matrix. There are no restrictions placed on  $S$  other than the requirement that  $|S| > r$  so the GMM estimate is well-defined. In particular,  $S$  may *exclude* some or all of the valid moment conditions contained in  $g$ . While this may seem strange, it accommodates a wider range of examples, including choosing between OLS and TSLS estimators.

To consider the limit distribution of  $\widehat{\theta}_S$ , we require some further notation. First define the derivative matrices

$$G = E[\nabla_{\theta} g(Z, \theta_0)], \quad H = E[\nabla_{\theta} h(Z, \theta_0)], \quad F = (G', H)'$$

and let  $\Omega = \text{Var}[f(Z, \theta_0)]$  where  $\Omega$  is partitioned into blocks  $\Omega_{gg}$ ,  $\Omega_{gh}$ ,  $\Omega_{hg}$ , and  $\Omega_{hh}$

---

<sup>3</sup>Identifying  $\tau$  requires further assumptions, as discussed in Section 2.3.

conformably with the partition of  $f$  by  $g$  and  $h$ . Notice that each of these expressions involves the *limiting random variable*  $Z$  rather than  $Z_{ni}$ , so that the corresponding expectations are taken with respect to a distribution for which all moment conditions are correctly specified. Finally, to avoid repeatedly writing out pre- and post-multiplication by  $\Xi_S$ , define  $F_S = \Xi_S F$  and  $\Omega_S = \Xi_S \Omega \Xi'_S$ . The following high level assumptions are sufficient for the consistency and asymptotic normality of the candidate GMM estimator  $\widehat{\theta}_S$ .

**Assumption 2.2** (High Level Sufficient Conditions).

- (a)  $\theta_0$  lies in the interior of  $\Theta$ , a compact set
- (b)  $\widetilde{W}_S \rightarrow_p W_S$ , a positive definite matrix
- (c)  $W_S \Xi_S E[f(Z, \theta)] = 0$  if and only if  $\theta = \theta_0$
- (d)  $E[f(Z, \theta)]$  is continuous on  $\Theta$
- (e)  $\sup_{\theta \in \Theta} \|f_n(\theta) - E[f(Z, \theta)]\| \rightarrow_p 0$
- (f)  $f$  is  $Z$ -almost surely differentiable in an open neighborhood  $\mathcal{B}$  of  $\theta_0$
- (g)  $\sup_{\theta \in \Theta} \|\nabla_{\theta} f_n(\theta) - F(\theta)\| \rightarrow_p 0$
- (h)  $\sqrt{n} f_n(\theta_0) \rightarrow_d M + \begin{bmatrix} 0 \\ \tau \end{bmatrix}$  where  $M \sim N_{p+q}(0, \Omega)$
- (i)  $F'_S W_S F_S$  is invertible

Although Assumption 2.2 closely approximates the standard regularity conditions for GMM estimation, establishing primitive conditions for Assumptions 2.2 (d), (e), (g) and (h) is slightly more involved under local mis-specification. Low-level sufficient conditions for the two running examples considered in this paper appear in Sections 2.4 and 2.5 below. For more general results, see Andrews (1988) Theorem 2 and Andrews (1992) Theorem 4. Notice that identification, (c), and continuity, (d), are conditions on the distribution of  $Z$ , the marginal law to which each  $Z_{ni}$  converges.

**Theorem 2.1** (Consistency). *Under Assumptions 2.1 and 2.2 (a)–(e),  $\widehat{\theta}_S \rightarrow_p \theta_0$ .*

**Theorem 2.2** (Asymptotic Normality). *Under Assumptions 2.1 and 2.2*

$$\sqrt{n}(\widehat{\theta}_S - \theta_0) \rightarrow_d -K_S \Xi_S \left( \begin{bmatrix} M_g \\ M_h \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

where  $K_S = [F'_S W_S F_S]^{-1} F'_S W_S$ ,  $M = (M'_g, M'_h)'$ , and  $M \sim N(0, \Omega)$ .



As we see from Theorems 2.1 and 2.2, *any* candidate GMM estimator  $\widehat{\theta}_S$  is consistent for  $\theta_0$  under local mis-specification. Unless  $S$  excludes *all* of the moment conditions contained in  $h$ , however,  $\widehat{\theta}_S$  inherits an asymptotic bias from the mis-specification parameter  $\tau$ . The local mis-specification framework is useful precisely because it results in a limit distribution for  $\widehat{\theta}_S$  with both a bias *and* a variance. This captures in asymptotic form the bias-variance tradeoff that we see in finite sample simulations. In contrast, fixed mis-specification results in a degenerate bias-variance tradeoff in the limit: scaling up by  $\sqrt{n}$  to yield an asymptotic variance causes the bias component to diverge.

### 2.3 Identification

Any form of moment selection requires an identifying assumption: we need to make clear which parameter value  $\theta_0$  counts as the “truth.” One approach, following Andrews (1999), is to assume that there exists a unique, maximal set of correctly specified moment conditions that identifies  $\theta_0$ . In the notation of the present paper<sup>4</sup> this is equivalent to the following:

**Assumption 2.3** (Andrews (1999) Identification Condition). *There exists a subset  $S_{max}$  of at least  $r$  moment conditions satisfying:*

(a)  $\Xi_{S_{max}} E[f(Z_{ni}, \theta_0)] = 0$

(b) *For any  $S' \neq S_{max}$  such that  $\Xi_{S'} E[f(Z_{ni}, \theta')] = 0$  for some  $\theta' \in \Theta$ ,  $|S_{max}| > |S'|$ .*

Andrews and Lu (2001) and Hong et al. (2003) take the same basic approach to identification, with appropriate modifications to allow for simultaneous model and moment selection. An advantage of Assumption 2.3 is that, under fixed mis-specification, it allows consistent selection of  $S_{max}$  without any prior knowledge of *which* moment conditions are correct. In the notation of the present paper this corresponds to having no moment conditions in the  $g$  block. As Hall (2005, p. 254) points out, however, the second part of Assumption 2.3 can fail even in simple settings such as linear instrumental variables estimation. When it does fail, the selected GMM estimator may no longer be consistent for  $\theta_0$ .

A different approach to identification is to assume that there is a minimal set of at least  $r$  moment conditions *known* to be correctly specified. This is the approach I follow here, as do Liao (2013) and Cheng and Liao (2013). With the exception of a short digression in Section 3.2, I maintain the following assumption throughout.

**Assumption 2.4** (FMSC Identification Condition). *Let  $\widehat{\theta}_v$  denote the GMM estimator based solely on the moment conditions contained in the  $g$ -block*

$$\widehat{\theta}_v = \arg \min_{\theta \in \Theta} g_n(\theta)' \widetilde{W}_v g_n(\theta)$$

---

<sup>4</sup>Although Andrews (1999), Andrews and Lu (2001), and Hong et al. (2003) consider *fixed* mis-specification, we can view this as a version of local mis-specification in which  $\tau \rightarrow \infty$ .

We call this the “valid estimator” and assume that it is identified, i.e. that  $p \geq r$ .

Assumption 2.4 and Theorem 2.2 immediately imply that the valid estimator shows no asymptotic bias.

**Corollary 2.1** (Limit Distribution of Valid Estimator). *Let  $S_v$  include only the moment conditions contained in  $g$ . Then, under Assumption 2.4 we have*

$$\sqrt{n}(\hat{\theta}_v - \theta_0) \rightarrow_d -K_v M_g$$

by applying Theorem 2.2 to  $S_v$ , where  $K_v = [G'W_v G]^{-1}G'W_v$  and  $M_g \sim N(0, \Omega_{gg})$ .

Both Assumptions 2.3 and 2.4 are strong, and neither fully nests the other. Which should be preferred is both a matter of taste and of the particular application one has in mind. In the context of the present paper, Assumption 2.4 is meant to represent the following situation. The  $g$ -block of moment conditions contains the “baseline” assumptions that an empirical researcher would be prepared to defend in a seminar while the  $h$ -block contains a set of stronger, more controversial “suspect” assumptions. These suspect are not wildly implausible – otherwise it would be strange to entertain them – but they are at least somewhat dubious. More than likely, they are not exactly true. In the FMSC framework, moment selection is carried out *conditional* on the baseline assumptions: we use the  $g$ -block to help us decide whether to include any of the moment conditions contained in the  $h$ -block. But why would we ever consider using controversial assumptions? The FMSC is designed for settings in which the  $h$ -block is expected to contain a substantial amount of information beyond that already contained in the  $g$ -block. The idea is that, if we knew the  $h$ -block was correctly specified, we would expect a large gain in efficiency by including it in estimation. This motivates the idea of trading off the variance reduction from including  $h$  against the potential increase in bias. Not all applications have the structure, but many do. I now consider two simple, empirically relevant examples.

## 2.4 Example: OLS versus TSLS

The simplest interesting application of the FMSC is choosing between ordinary least squares (OLS) and two-stage least squares (TSLS) estimators of the effect  $\beta$  of a single endogenous regressor  $x$  on an outcome of interest  $y$ . To keep the presentation transparent, I work within an iid, homoskedastic setting. Neither of these restrictions, however, is necessary. Without loss of generality assume that there are no exogenous regressors, or equivalently that they have been “projected out of the system” so that the data generating process is given by

$$y_i = \beta x_i + \epsilon_i \tag{1}$$

$$x_i = \mathbf{z}'_i \boldsymbol{\pi} + v_i \tag{2}$$

where  $\beta$  and  $\boldsymbol{\pi}$  are unknown constants,  $\mathbf{z}_i$  is a vector of exogenous and relevant instruments,  $x_i$  is the endogenous regressor,  $y_i$  is the outcome of interest, and  $\epsilon_i, v_i$  are unobservable error terms. All random variables in this system are mean zero, or equivalently all constant terms have been projected out. Stacking observations in the usual way, let  $\mathbf{z}' = (z_1, \dots, z_n)$ ,  $Z' = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ ,  $\mathbf{x}' = (x_1, \dots, x_n)$  and so on. The two estimators under consideration are  $\widehat{\beta}_{OLS}$  and  $\widetilde{\beta}_{TSLs}$ , given by

$$\widehat{\beta}_{OLS} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \quad (3)$$

$$\widetilde{\beta}_{TSLs} = (\mathbf{x}'P_Z\mathbf{x})^{-1}\mathbf{x}'P_Z\mathbf{y} \quad (4)$$

where  $P_Z = Z(Z'Z)^{-1}Z'$ . But why should we ever *consider* using OLS if  $x$  is endogenous and we have valid instruments at our disposal? The answer is simple: TSLs is a high variance estimator. Depending on the degree of endogeneity present in  $x$  and the strength of the instruments,  $\widehat{\beta}_{OLS}$  could easily have the lower mean-squared error of the two estimators.<sup>5</sup> The appropriate version of the local mis-specification assumption for this example is as follows

$$E \begin{bmatrix} \mathbf{z}_{ni}\epsilon_{ni} \\ x_{ni}\epsilon_{ni} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \tau/\sqrt{n} \end{bmatrix} \quad (5)$$

where  $\mathbf{0}$  is a vector of zeros, and  $\tau$  is an unknown constant that encodes the endogeneity of  $x$ . In this setting, there is only a single moment condition in the  $h$ -block, namely  $E[x_{ni}\epsilon_{ni}] = \tau/\sqrt{n}$ . The question is not whether we should use this moment condition *in addition* to the TSLs moment conditions written above it, but rather whether we should use it *instead* of them. The following simple low-level conditions are sufficient for the asymptotic normality of the OLS and TSLs estimators in this example.

**Assumption 2.5** (OLS versus TSLs). *Let  $\{(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni}) : 1 \leq i \leq n, n = 1, 2, \dots\}$  be a triangular array of random variables such that*

- (a)  $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni}) \sim iid$  and mean zero within each row of the array (i.e. for fixed  $n$ )
- (b)  $E[\mathbf{z}_{ni}\epsilon_{ni}] = \mathbf{0}$ ,  $E[\mathbf{z}_{ni}v_{ni}] = \mathbf{0}$ , and  $E[\epsilon_{ni}v_{ni}] = \tau/\sqrt{n}$  for all  $n$
- (c)  $E[|\mathbf{z}_{ni}|^{4+\eta}] < C$ ,  $E[|\epsilon_{ni}|^{4+\eta}] < C$ , and  $E[|v_{ni}|^{4+\eta}] < C$  for some  $\eta > 0$ ,  $C < \infty$
- (d)  $E[\mathbf{z}_{ni}\mathbf{z}'_{ni}] \rightarrow Q > 0$ ,  $E[v_{ni}^2] \rightarrow \sigma_v^2 > 0$ , and  $E[\epsilon_{ni}^2] \rightarrow \sigma_\epsilon^2 > 0$  as  $n \rightarrow \infty$
- (e) As  $n \rightarrow \infty$ ,  $E[\epsilon_{ni}^2\mathbf{z}_{ni}\mathbf{z}'_{ni}] - E[\epsilon_{ni}^2]E[\mathbf{z}_{ni}\mathbf{z}'_{ni}] \rightarrow 0$ ,  $E[\epsilon_{ni}^2v_{ni}\mathbf{z}'_{ni}] - E[\epsilon_{ni}^2]E[v_{ni}\mathbf{z}'_{ni}] \rightarrow 0$ , and  $E[\epsilon_{ni}^2v_{ni}^2] - E[\epsilon_{ni}^2]E[v_{ni}^2] \rightarrow 0$
- (f)  $x_{ni} = \mathbf{z}'_{ni}\boldsymbol{\pi} + v_i$  where  $\boldsymbol{\pi} \neq \mathbf{0}$ , and  $y_{ni} = \beta x_{ni} + \epsilon_{ni}$

---

<sup>5</sup>Because the moments of the TSLs estimator only exist up to the order of overidentification (Phillips, 1980) this statement should be understood to refer to “trimmed” mean-squared error when the number of instruments is two or fewer. See, e.g., Hansen (2013).

Parts (a), (b) and (d) correspond to the local mis-specification assumption, part (c) is a set of moment restrictions, and (f) is simply the DGP. Part (e) is the homoskedasticity assumption: an *asymptotic* restriction on the joint distribution of  $v_{ni}$ ,  $\epsilon_{ni}$ , and  $\mathbf{z}_{ni}$ . This condition holds automatically, given the other assumptions, if  $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni})$  are jointly normal, as in our simulation experiment described below.

**Theorem 2.3** (OLS and TSLS Limit Distributions). *Under Assumption 2.5,*

$$\begin{bmatrix} \sqrt{n}(\widehat{\beta}_{OLS} - \beta) \\ \sqrt{n}(\widehat{\beta}_{TSLS} - \beta) \end{bmatrix} \xrightarrow{d} N \left( \begin{bmatrix} \tau/\sigma_x^2 \\ 0 \end{bmatrix}, \sigma_\epsilon^2 \begin{bmatrix} 1/\sigma_x^2 & 1/\sigma_x^2 \\ 1/\sigma_x^2 & 1/\gamma^2 \end{bmatrix} \right)$$

where  $\sigma_x^2 = \gamma^2 + \sigma_v^2$ ,  $\gamma^2 = \boldsymbol{\pi}'Q\boldsymbol{\pi}$ , and  $Q$ ,  $\sigma_\epsilon^2$ , and  $\sigma_v^2$  are defined in Assumption 2.5.

We see from the preceding result that the variance of the OLS estimator is always strictly lower than that of the TSLS estimator since  $\sigma_\epsilon^2/\sigma_x^2 = \sigma_\epsilon^2/(\gamma^2 + \sigma_v^2)$ . Unless  $\tau = 0$ , however, OLS shows an asymptotic bias. In contrast, the TSLS estimator is asymptotically unbiased regardless of the value of  $\tau$ .

## 2.5 Example: Choosing Instrumental Variables

The preceding example was quite specific, but in a sense it amounted to a problem of instrument selection: if  $x$  is exogenous, it is clearly “its own best instrument.” Viewed from this perspective, the FMSC amounted to trading off endogeneity against instrument strength. I now consider instrument selection *in general* for linear GMM estimators in an iid setting. Consider the following model:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i \tag{6}$$

$$\mathbf{x}_i = \Pi_1' \mathbf{z}_i^{(1)} + \Pi_2' \mathbf{z}_i^{(2)} + \mathbf{v}_i \tag{7}$$

where  $y$  is an outcome of interest,  $\mathbf{x}$  is an  $r$ -vector of regressors, some of which are endogenous,  $\mathbf{z}^{(1)}$  is a  $p$ -vector of instruments known to be exogenous, and  $\mathbf{z}^{(2)}$  is a  $q$ -vector of *potentially endogenous* instruments. The  $r$ -vector  $\boldsymbol{\beta}$ ,  $p \times r$  matrix  $\Pi_1$ , and  $q \times r$  matrix  $\Pi_2$  contain unknown constants. Stacking observations in the usual way, let  $\mathbf{y}' = (y_1, \dots, y_n)$ ,  $X' = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ ,  $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ ,  $Z_1' = (\mathbf{z}_1^{(1)}, \dots, \mathbf{z}_n^{(1)})$ ,  $Z_2' = (\mathbf{z}_1^{(2)}, \dots, \mathbf{z}_n^{(2)})$ , and  $V' = (\mathbf{v}_1, \dots, \mathbf{v}_n)$  where  $n$  is the sample size. Using this notation,  $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$  and  $X = Z\Pi + V$ , where  $Z = (Z_1, Z_2)$  and  $\Pi = (\Pi_1', \Pi_2')'$ .

The idea behind this setup is that the instruments contained in  $Z_2$  are expected to be strong. If we were confident that they were exogenous, we would certainly use them in estimation. Yet the very fact that we expect them to be strongly correlated with  $\mathbf{x}$  gives us reason to fear that the instruments contained in  $Z_2$  may be endogenous. The exact opposite is true of  $Z_1$ . These are the instruments that we are prepared to assume are exogenous. But when is such an assumption plausible? Precisely when the instruments contained in  $Z_1$  are *not especially strong*. In this setting, the FMSC attempts to trade

off a small increase in bias from using a *slightly* endogenous instrument against a larger decrease in variance from increasing the overall strength of the instruments used in estimation.

To this end, consider a general linear GMM estimator of the form

$$\widehat{\beta}_S = (X'Z_S\widetilde{W}_SZ_S'X)^{-1}X'Z_S\widetilde{W}_SZ_S'\mathbf{y}$$

where  $S$  indexes the instruments used in estimation,  $Z_S' = \Xi_S Z'$  is the matrix containing only those instruments included in  $S$ ,  $|S|$  is the number of instruments used in estimation and  $\widetilde{W}_S$  is an  $|S| \times |S|$  positive semi-definite weighting matrix. In this example, the local mis-specification assumption is given by

$$E \begin{bmatrix} \mathbf{z}_{ni}^{(1)}(y_i - \mathbf{x}_i\beta) \\ \mathbf{z}_{ni}^{(2)}(y_i - \mathbf{x}_i\beta) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\tau}/\sqrt{n} \end{bmatrix} \quad (8)$$

where  $\mathbf{0}$  is a  $p$ -vector of zeros, and  $\boldsymbol{\tau}$  is a  $q$ -vector of unknown constants. The following low-level conditions are sufficient for the asymptotic normality of  $\widehat{\beta}_S$ .

**Assumption 2.6** (Choosing IVs). *Let  $\{(\mathbf{z}_{ni}, \mathbf{v}_{ni}, \epsilon_{ni}) : 1 \leq i \leq n, n = 1, 2, \dots\}$  be a triangular array of random variables with  $\mathbf{z}_{ni} = (\mathbf{z}_{ni}^{(1)}, \mathbf{z}_{ni}^{(2)})$  such that*

- (a)  $(\mathbf{z}_{ni}, \mathbf{v}_{ni}, \epsilon_{ni}) \sim iid$  within each row of the array (i.e. for fixed  $n$ )
- (b)  $E[\mathbf{v}_{ni}\mathbf{z}_{ni}'] = \mathbf{0}$ ,  $E[\mathbf{z}_{ni}^{(1)}\epsilon_{ni}] = \mathbf{0}$ , and  $E[\mathbf{z}_{ni}^{(2)}\epsilon_{ni}] = \boldsymbol{\tau}/\sqrt{n}$  for all  $n$
- (c)  $E[|\mathbf{z}_{ni}|^{4+\eta}] < C$ ,  $E[|\epsilon_{ni}|^{4+\eta}] < C$ , and  $E[|\mathbf{v}_{ni}|^{4+\eta}] < C$  for some  $\eta > 0$ ,  $C < \infty$
- (d)  $E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \rightarrow Q > 0$  and  $E[\epsilon_{ni}^2\mathbf{z}_{ni}\mathbf{z}_{ni}'] \rightarrow \Omega > 0$  as  $n \rightarrow \infty$
- (e)  $\mathbf{x}_{ni} = \Pi_1'\mathbf{z}_{ni}^{(1)} + \Pi_2'\mathbf{z}_{ni}^{(2)} + \mathbf{v}_{ni}$  where  $\Pi_1 \neq \mathbf{0}$ ,  $\Pi_2 \neq \mathbf{0}$ , and  $y_i = \mathbf{x}_{ni}'\beta + \epsilon_{ni}$

The preceding conditions are similar to although more general than those contained in Assumption 2.5. Although I no longer assume homoskedasticity, for simplicity I retain the assumption that the triangular array is iid in each row.

**Theorem 2.4** (Choosing IVs Limit Distribution). *Suppose that  $\widetilde{W}_S \rightarrow_p W_S > 0$ . Then, under Assumption 2.6*

$$\sqrt{n}(\widehat{\beta}_S - \beta) \xrightarrow{d} -K_S\Xi_S \left( \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\tau} \end{bmatrix} + M \right)$$

where

$$-K_S = (\Pi'Q_S W_S Q_S' \Pi)^{-1} \Pi'Q_S W_S$$

$M \sim N(\mathbf{0}, \Omega)$ ,  $Q_S = Q\Xi_S'$ , and  $Q$  and  $\Omega$  are defined in Assumption 2.6.

### 3 The Focused Moment Selection Criterion

#### 3.1 The General Case

The FMSC chooses among the potentially invalid moment conditions contained in  $h$  to minimize estimator AMSE for a target parameter. Denote this target parameter by  $\mu$ , a real-valued,  $Z$ -almost continuous function of the parameter vector  $\theta$  that is differentiable in a neighborhood of  $\theta_0$ . Further, define the GMM estimator of  $\mu$  based on  $\hat{\theta}_S$  by  $\hat{\mu}_S = \mu(\hat{\theta}_S)$  and the true value of  $\mu$  by  $\mu_0 = \mu(\theta_0)$ . Applying the Delta Method to Theorem 2.2 gives the AMSE of  $\hat{\mu}_S$ .

**Corollary 3.1** (AMSE of Target Parameter). *Under the hypotheses of Theorem 2.2,*

$$\sqrt{n}(\hat{\mu}_S - \mu_0) \rightarrow_d -\nabla_{\theta}\mu(\theta_0)'K_S\Xi_S \left( M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

where  $M$  is defined in Theorem 2.2. Hence,

$$AMSE(\hat{\mu}_S) = \nabla_{\theta}\mu(\theta_0)'K_S\Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \tau\tau' \end{bmatrix} + \Omega \right\} \Xi_S'K_S'\nabla_{\theta}\mu(\theta_0).$$

For the valid estimator  $\hat{\theta}_v$  we have  $K_v = [G'W_vG]^{-1}G'W_v$  and  $\Xi_v = [\mathbf{I}_p \quad \mathbf{0}_{p \times q}]$ . Thus, the valid estimator  $\hat{\mu}_v$  of  $\mu$  has zero asymptotic bias. In contrast, any candidate estimator  $\hat{\mu}_S$  that includes moment conditions from  $h$  inherits an asymptotic bias from the corresponding elements of  $\tau$ , the extend and direction of which depends both on  $K_S$  and  $\nabla_{\theta}\mu(\theta_0)$ . Adding moment conditions from  $h$ , however, generally decreases asymptotic variance. In particular, the usual proof that adding moment conditions cannot increase asymptotic variance under efficient GMM (see for example Hall, 2005, ch. 6) continues to hold under local mis-specification, because all moment conditions are correctly specified in the limit. Thus, we see that local mis-specification gives an asymptotic analogue of the bias-variance tradeoff that we encounter in finite samples.<sup>6</sup>

Using this framework for moment selection requires estimators of the unknown quantities:  $\theta_0$ ,  $K_S$ ,  $\Omega$ , and  $\tau$ . Under local mis-specification, the estimator of  $\theta$  under *any* moment set is consistent. A natural estimator is  $\hat{\theta}_v$ , although there are other possibilities. Recall that  $K_S = [F_S'W_SF_S]^{-1}F_S'W_S\Xi_S$ . Because it is simply the selection matrix defining moment set  $S$ ,  $\Xi_S$  is known. The remaining quantities  $F_S$  and  $W_S$  that make up  $K_S$  are consistently estimated by their sample analogues under Assumption 2.2. Similarly, consistent estimators of  $\Omega$  are readily available under local mis-specification, although the precise form depends on the situation.<sup>7</sup> The only

<sup>6</sup>The general result for adding moment conditions in GMM is only relevant in situations where the valid moment set is strictly nested inside of all other candidate moment sets. When this does not hold, such as in the OLS versus IV example, we establish an analogous ordering of asymptotic variances by direct calculation.

<sup>7</sup>See Sections 3.3 and 3.5 for discussion of this point for the two running examples.

remaining unknown is  $\tau$ . Local mis-specification is essential for making meaningful comparisons of AMSE because it prevents the bias term from dominating the comparison. Unfortunately, it also prevents consistent estimation of the asymptotic bias parameter. Under Assumption 2.4, however, it remains possible to construct an *asymptotically unbiased* estimator  $\hat{\tau}$  of  $\tau$  by substituting  $\hat{\theta}_v$ , the estimator of  $\theta_0$  that uses only correctly specified moment conditions, into  $h_n$ , the sample analogue of the potentially mis-specified moment conditions. In other words,  $\hat{\tau} = \sqrt{n}h_n(\hat{\theta}_v)$ .

**Theorem 3.1** (Asymptotic Distribution of  $\hat{\tau}$ ). *Let  $\hat{\tau} = \sqrt{n}h_n(\hat{\theta}_v)$  where  $\hat{\theta}_v$  is the valid estimator, based only on the moment conditions contained in  $g$ . Then under Assumptions 2.1, 2.2 and 2.4*

$$\hat{\tau} \rightarrow_d \Psi \left( M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right), \quad \Psi = \begin{bmatrix} -HK_v & \mathbf{I}_q \end{bmatrix}$$

where  $K_v$  is defined in Corollary 2.1. Thus,  $\hat{\tau} \rightarrow_d (\Psi M + \tau) \sim N_q(\tau, \Psi\Omega\Psi')$ .

Returning to Corollary 3.1, however, we see that it is  $\tau\tau'$  rather than  $\tau$  that enters the expression for AMSE. Although  $\hat{\tau}$  is an asymptotically unbiased estimator of  $\tau$ , the limiting expectation of  $\hat{\tau}\hat{\tau}'$  is not  $\tau\tau'$  because  $\hat{\tau}$  has an asymptotic variance. Subtracting a consistent estimate of the asymptotic variance removes this asymptotic bias.

**Corollary 3.2** (Asymptotically Unbiased Estimator of  $\tau\tau'$ ). *If  $\hat{\Omega}$  and  $\hat{\Psi}$  are consistent for  $\Omega$  and  $\Psi$ , then  $\hat{\tau}\hat{\tau}' - \hat{\Psi}\hat{\Omega}\hat{\Psi}'$  is an asymptotically unbiased estimator of  $\tau\tau'$ .*

It follows that

$$\text{FMSC}_n(S) = \nabla_{\theta}\mu(\hat{\theta})' \hat{K}_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \hat{\tau}\hat{\tau}' - \hat{\Psi}\hat{\Omega}\hat{\Psi}' \end{bmatrix} + \hat{\Omega} \right\} \Xi_S' \hat{K}_S' \nabla_{\theta}\mu(\hat{\theta}) \quad (9)$$

provides an asymptotically unbiased estimator of AMSE. Given a set  $\mathcal{S}$  of candidate specifications, the FMSC selects the candidate  $S^*$  that *minimizes* the expression given in Equation 9, that is  $S_{FMSC}^* = \arg \min_{S \in \mathcal{S}} \text{FMSC}_n(S)$ .

At this point, it is worth taking a brief pause to survey the ground covered thus far. We began with a target parameter,  $\mu$ , a risk function, mean-squared error, and a collection of candidate estimators,  $\hat{\mu}_S$  for  $S \in \mathcal{S}$ . Our goal was to choose the estimator with the lowest risk. Because finite-sample distributions were unavailable, we resorted to an asymptotic experiment, local mis-specification, that preserved the bias-variance tradeoff embodied in our chosen risk function. We then calculated the risk of the *limit distribution* of  $\hat{\mu}_S$  to use as a stand-in for the finite-sample risk. This quantity involved several unknown parameters. We estimated these in such a way that the resulting asymptotic risk estimate would converge in distribution to a random variable with mean equal to the true asymptotic risk. The result was the FMSC: an asymptotically unbiased estimator of the AMSE of  $\hat{\mu}_S$ . Viewing the FMSC at this level of abstraction raises two questions. First, could we have chosen a risk function other

than mean-squared error? Second, why should we use an *asymptotically unbiased* risk estimator?

The answer to the first question is a straightforward yes. The idea of using asymptotic risk as a stand-in for finite sample risk requires only that we can characterize the limit distribution of each  $\hat{\mu}_S$  and use it to evaluate the chosen risk function. [Claeskens et al. \(2006\)](#) and [Claeskens and Hjort \(2008a\)](#), for example, show how the FIC for model selection in maximum likelihood models can be extended from squared error to  $L_p$  and linex loss, respectively, in precisely this way. One could easily do the same for the FMSC although I do not consider this possibility further here.

Answering the second question is more difficult. Under local mis-specification it is impossible to consistently estimate AMSE.<sup>8</sup> If we merely use the plug-in estimator of the squared asymptotic bias based on  $\hat{\tau}$ , the resulting AMSE estimate will “overshoot” asymptotically. Accordingly, it seems natural to correct this bias as explained in [Corollary 3.2](#). This is the same heuristic that underlies the classical AIC and TIC model selection criteria as well as more recent procedures such as those described in [Claeskens and Hjort \(2003\)](#) and [Schorfheide \(2005\)](#). A heuristic, however, is not a rigorous justification: there could certainly be situations in which it makes sense to use a risk estimator other than the asymptotically unbiased one suggested here. If one wished to consider risk functions other than MSE, to take a simple example, it may not be possible to derive an asymptotically unbiased risk estimator. The plug-in estimator, however, is always available. Although I do not consider them further below, alternative risk estimators in the FMSC framework could be an interesting topic for future research.

### 3.2 Digression: Failure of the Identification Condition

When  $r > p$ , [Assumption 2.4](#) fails:  $\theta_0$  is not estimable by  $\hat{\theta}_v$ , so  $\hat{\tau}$  is an infeasible estimator of  $\tau$ . A naïve approach to this problem would be to substitute another consistent estimator of  $\theta_0$  and proceed analogously. Unfortunately, this approach fails. To understand why, consider the case in which all moment conditions are potentially invalid so that the  $g$ -block is empty. Letting  $\hat{\theta}_f$  denote the estimator based on the full set of moment conditions in  $h$ ,  $\sqrt{n}h_n(\hat{\theta}_f) \rightarrow_d \Gamma\mathcal{N}_q(\tau, \Omega)$  where  $\Gamma = \mathbf{I}_q - H(H'WH)^{-1}H'W$ , using an argument similar to that in the proof of [Theorem 3.1](#). The mean,  $\Gamma\tau$ , of the resulting limit distribution does not equal  $\tau$ , and because  $\Gamma$  has rank  $q - r$  we cannot pre-multiply by its inverse to extract an estimate of  $\tau$ . Intuitively,  $q - r$  over-identifying restrictions are insufficient to estimate a  $q$ -vector:  $\tau$  is not identified unless we have a minimum of  $r$  valid moment conditions. However, the limiting distribution of  $\sqrt{n}h_n(\hat{\theta}_f)$  partially identifies  $\tau$  even when we have no valid moment conditions at our disposal. A combination of this information with prior restrictions on the magnitude of the components of  $\tau$  allows the use of the FMSC framework to carry out a sensitivity analysis

---

<sup>8</sup>This is not a defect of the FMSC: there is a fundamental trade-off between consistency and desirable risk properties. See [Section 4](#) for a discussion of this point.



when  $r > p$ . For example, the worst-case estimate of AMSE over values of  $\tau$  in the identified region could still allow certain moment sets to be ruled out. This idea shares similarities with [Kraay \(2012\)](#) and [Conley et al. \(2012\)](#), two recent papers that suggest methods for evaluating the robustness of conclusions drawn from IV regressions when the instruments used may be invalid.

### 3.3 FMSC for OLS versus TSLS Example

The FMSC has a particularly convenient and transparent form in the OLS versus TSLS example introduced in Section 2.4. Since the target parameter in this case is simply  $\beta$ , the FMSC amounts to comparing the AMSE of OLS to that of TSLS. As an immediate consequence of Theorem 2.3,

$$\text{AMSE(OLS)} = \frac{\tau^2}{\sigma_x^4} + \frac{\sigma_\epsilon^2}{\sigma_x^2}, \quad \text{AMSE(TSLS)} = \frac{\sigma_\epsilon^2}{\gamma^2}.$$

Rearranging, we see that the AMSE of the OLS estimator is strictly less than that of the TSLS estimator whenever  $\tau^2 < \sigma_x^2 \sigma_\epsilon^2 \sigma_v^2 / \gamma^2$ . Using this expression for moment selection, of course, requires estimates of the unknown parameters. Fortunately, the familiar estimators of  $\sigma_x^2$ ,  $\gamma^2$ , and  $\sigma_v^2$  remain consistent under Assumption 2.5. I set:

$$\hat{\sigma}_x^2 = n^{-1} \mathbf{x}' \mathbf{x}, \quad \hat{\gamma}^2 = n^{-1} \mathbf{x}' Z (Z' Z)^{-1} Z' \mathbf{x}, \quad \hat{\sigma}_v^2 = \hat{\sigma}_x^2 - \hat{\gamma}^2.$$

To estimate  $\sigma_\epsilon^2$  there are two choices: one can either use the residuals from the OLS estimator or those from the TSLS estimator. Under local mis-specification, both provide consistent estimators of  $\sigma_\epsilon^2$ . Of the two, the estimator based on the TSLS residuals should be more robust, however, unless the instruments are quite weak. This is because the exogeneity of  $x$ , even though it disappears in the limit under our asymptotics, is non-zero in finite samples. Thus, I use

$$\hat{\sigma}_\epsilon^2 = n^{-1} \left( \mathbf{y} - \mathbf{x} \tilde{\beta}_{TSLS} \right)' \left( \mathbf{y} - \mathbf{x} \tilde{\beta}_{TSLS} \right)$$

to estimate  $\sigma_\epsilon^2$  below. All that remains is to estimate  $\tau^2$ . Specializing Theorem 3.1 and Corollary 3.2 to the present example gives the following result.

**Theorem 3.2.** *Let  $\hat{\tau} = n^{-1/2} \mathbf{x}' (\mathbf{y} - \mathbf{x} \tilde{\beta}_{TSLS})$ . Under Assumption 2.5 we have*

$$\hat{\tau} \rightarrow_d N(\tau, V), \quad V = \sigma_\epsilon^2 \sigma_x^2 (\sigma_v^2 / \gamma^2).$$

It follows that  $\hat{\tau}^2 - \hat{\sigma}_\epsilon^2 \hat{\sigma}_x^2 (\hat{\sigma}_v^2 / \hat{\gamma}^2)$  is an asymptotically unbiased estimator of  $\tau^2$  and hence, substituting into the AMSE inequality from above and rearranging, the FMSC instructs us to choose OLS whenever  $\hat{T}_{FMSC} = \hat{\tau}^2 / \hat{V} < 2$  where  $\hat{V} = \hat{\sigma}_\epsilon^2 \hat{\sigma}_x^2 \hat{\sigma}_v^2 / \hat{\gamma}^2$ . The quantity  $\hat{T}_{FMSC}$  looks very much like a test statistic and indeed it can be viewed as such. By Theorem 3.2 and the continuous mapping theorem,  $\hat{T}_{FMSC} \rightarrow_d \chi^2(1)$ . Thus,

the FMSC can be viewed as a test of the null hypothesis  $H_0: \tau = 0$  against the two-sided alternative with a critical value of 2. This corresponds to a significance level of  $\alpha \approx 0.16$ .

But how does this novel “test” compare to something more familiar, say the Durbin-Hausman-Wu (DHW) test? It turns out that in this particular example, although not in general, carrying out moment selection via the FMSC is *numerically equivalent* to using OLS unless the DHW test rejects at the 16% level. In other words,  $\widehat{T}_{FMSC} = \widehat{T}_{DHW}$ . To see why this is so first note that

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLs} \\ \widetilde{\beta}_{TSLs} - \beta \end{pmatrix} = \begin{bmatrix} 1 & -1 \end{bmatrix} \sqrt{n} \begin{pmatrix} \widehat{\beta}_{OLS} - \beta \\ \widetilde{\beta}_{TSLs} - \beta \end{pmatrix} \rightarrow_d N(\tau/\sigma_x^2, \Sigma)$$

by Theorem 3.2, where

$$\Sigma = \text{AVAR}(TSLs) - \text{AVAR}(OLS) = \sigma_\epsilon^2 (1/\gamma^2 - 1/\sigma_x^2).$$

Thus, under  $H_0: \tau = 0$ , the DHW test statistic

$$\widehat{T}_{DHW} = n \widehat{\Sigma}^{-1} (\widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLs})^2 = \frac{n(\widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLs})^2}{\widehat{\sigma}_\epsilon^2 (1/\widehat{\gamma}^2 - 1/\widehat{\sigma}_x^2)}$$

converges in distribution to a  $\chi^2(1)$  random variable. Now, rewriting  $\widehat{V}$ , we find that

$$\widehat{V} = \widehat{\sigma}_\epsilon^2 \widehat{\sigma}_x^2 \left( \frac{\widehat{\sigma}_v^2}{\widehat{\gamma}^2} \right) = \widehat{\sigma}_\epsilon^2 \widehat{\sigma}_x^2 \left( \frac{\widehat{\sigma}_x^2 - \widehat{\gamma}^2}{\widehat{\gamma}^2} \right) = \widehat{\sigma}_\epsilon^2 \widehat{\sigma}_x^4 \left( \frac{1}{\widehat{\gamma}^2} - \frac{1}{\widehat{\sigma}_x^2} \right) = \widehat{\sigma}_x^4 \widehat{\Sigma}$$

using the fact that  $\widehat{\sigma}_v = \widehat{\sigma}_x^2 - \widehat{\gamma}^2$ . Thus, to show that  $\widehat{T}_{FMSC} = \widehat{T}_{DHW}$ , all that remains is to establish that  $\widehat{\tau}^2 = n \widehat{\sigma}_x^4 (\widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLs})^2$ , which we obtain as follows:

$$\widehat{\tau}^2 = \left[ n^{-1/2} \mathbf{x}'(\mathbf{y} - \mathbf{x}\widetilde{\beta}) \right]^2 = n^{-1} \left[ \mathbf{x}'\mathbf{x} (\widehat{\beta} - \widetilde{\beta}) \right]^2 = n^{-1} \left[ n \widehat{\sigma}_x^2 (\widehat{\beta} - \widetilde{\beta}) \right]^2.$$

The equivalence between FMSC selection and a DHW test in this example is helpful for two reasons. First, it provides a novel justification for the use of the DHW test to select between OLS and TSLs. So long as it is carried out with  $\alpha \approx 16\%$ , the DHW test is equivalent to selecting the estimator that minimizes an asymptotically unbiased estimator of AMSE. Note that this significance level differs from the more usual values of 5% or 10% in that it leads us to select TSLs *more often*: OLS should indeed be given the benefit of the doubt, but not by so wide a margin as traditional practice suggests. Second, this equivalence shows that the FMSC can be viewed as an *extension* of the machinery behind the familiar DHW test to more general GMM environments. Naturally each application of the FMSC should be evaluated on its own merits, but it is reassuring that the local mis-specification framework leads to a reasonable procedure in a simple setting where it can be compared to more familiar

techniques.

### 3.4 Simulation Study: OLS versus TSLS Example

I now examine the performance of the FMSC when used to choose between OLS and TSLS in a simple simulation experiment. All calculations in this section are based on the formulas from Sections 2.4 and 3.3 and 10,000 simulation replications.

The data generating process is given by

$$y_i = 0.5x_i + \epsilon_i \tag{10}$$

$$x_i = \pi(z_{1i} + z_{2i} + z_{3i}) + v_i \tag{11}$$

where  $(\epsilon_i, v_i, z_{1i}, z_{2i}, z_{3i}) \sim \text{iid } N(0, \mathcal{S})$

$$\mathcal{S} = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 - \pi^2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/3 \end{bmatrix} \tag{12}$$

for  $i = 1, \dots, N$  and we vary  $N$ ,  $\rho$  and  $\pi$  over a grid. The goal is to estimate the effect of  $x$  on  $y$ , in this case 0.5, with minimum MSE by choosing between OLS and TSLS estimators. To ensure that the finite-sample MSE of the TSLS estimator exists, this DGP includes three instruments leading to two overidentifying restrictions (Phillips, 1980).<sup>9</sup> This design satisfies the conditions of Assumption 2.5 and keeps the variance of  $x$  fixed at one so that  $\pi = \text{Corr}(x_i, z_{1i} + z_{2i} + z_{3i})$  and  $\rho = \text{Corr}(x_i, \epsilon_i)$ . The first-stage R-squared is simply  $1 - \sigma_v^2/\sigma_x^2 = \pi^2$  so that larger values of  $|\pi|$  reduce the variance of the TSLS estimator. Since  $\rho$  controls the endogeneity of  $x$ , larger values of  $|\rho|$  increase the bias of the OLS estimator.

Figure 1 compares the root mean-squared error (RMSE) of the post-FMSC estimator to those of the simple OLS and TSLS estimators. For any values of  $N$  and  $\pi$  there is a value of  $\rho$  below which OLS outperforms TSLS. As  $N$  and  $\pi$  increase this value approaches zero; as they decrease it approaches one. Although the first two moments of the TSLS estimator exist in this simulation design, none of its higher moments do. This fact is clearly evident for small values of  $N$  and  $\pi$ : even with 10,000 simulation replications, the RMSE of the TSLS estimator shows a noticeable degree of simulation error unlike those of the OLS and post-FMSC estimators.

The FMSC represents a compromise between OLS and TSLS. When the RMSE of TSLS is high, the FMSC behaves more like OLS; when the RMSE of OLS is high it behaves more like TSLS. Because the FMSC is itself a random variable, however, it sometimes makes moment selection mistakes.<sup>10</sup> For this reason it does not attain an

<sup>9</sup>Alternatively, one could use fewer instruments in the DGP and use work with trimmed MSE.

<sup>10</sup>For more discussion of this point, see Section 4.

RMSE equal to the lower envelope of the OLS and TSLS estimators: moment selection is not a free lunch. The larger the RMSE difference between OLS and TSLS, however, the closer the FMSC comes to this lower envelope: the more costly the mistake, the rarer.

The question remains: should one carry out moment selection via the FMSC or avoid selection altogether and simply use the correctly specified TSLS estimator? As usual the answer depends where we happen to be in the parameter space. Because it is sometimes makes mistakes, the post-FMSC estimator cannot provide a uniform improvement over TSLS. And indeed, as seen from the simulation results, it does not. The worst-case RMSE of the post-FMSC estimator, however, is never much higher than that of TSLS while the worst-case RMSE of TSLS can be dramatically higher than that of the FMSC, depending on parameter values. Moreover, as discussed above, the FMSC is intended for situations in which an applied researcher fears that her “baseline” assumptions may be too weak and consequently considers adding one or more “controversial” assumptions. In this case, she fears that the exogenous instruments  $z_1, z_2, z_3$  are not particularly strong,  $\pi$  is small relative to  $N$ , and thus entertains the assumption that  $x$  is exogenous. It is precisely in these situations that the FMSC is likely to provide large performance gains over TSLS.

As shown in the preceding section, the FMSC takes a very special form in this example: it is equivalent to a DHW test with  $\alpha \approx 0.16$ . Accordingly, Figure 2 compares the RMSE of the post-FMSC estimator to those of DHW pre-test estimators with significance levels  $\alpha = 0.05$  and  $\alpha = 0.1$ , indicated in the legend by DHW95 and DHW90. Since these three procedures differ only in their critical values, they show similar qualitative behavior. When  $\rho$  is sufficiently close to zero, we saw from Figure 1 that OLS has a lower RMSE than TSLS. Accordingly since DHW95 and DHW90 require a higher burden of proof to reject OLS in favor of TSLS, they outperform FMSC in this region of the parameter space. When  $\rho$  crosses the threshold beyond which TSLS has a lower RMSE than OLS, the tables are turned: FMSC outperforms DHW95 and DHW90. As  $\rho$  increases further, relative to sample size and  $\pi$ , the three procedures become indistinguishable in terms of RMSE. There is nothing mysterious about this behavior: it is merely a consequence of the fact that DHW95 employs the largest critical value, followed by DHW90, while FMSC employs the smallest. In practice, of course,  $\rho$  is unknown so we cannot use it to decide between DHW95, DHW90 and FMSC. In terms of worst-case RMSE, however, FMSC gives the best performance of the three, particularly for larger sample sizes.

### 3.5 FMSC for Instrument Selection Example

As explained above, the OLS versus TSLS example can be viewed as a special case of the more general instrument selection problem from Section 2.5. To apply the FMSC to this problem, we simply need to specialize Equation 9. Recall that  $Z_1$  is the matrix containing the  $p$  instruments known to be exogenous,  $Z_2$  is the matrix containing the  $q$  potentially endogenous instruments and  $Z = (Z_1, Z_2)$ . To simplify the notation in

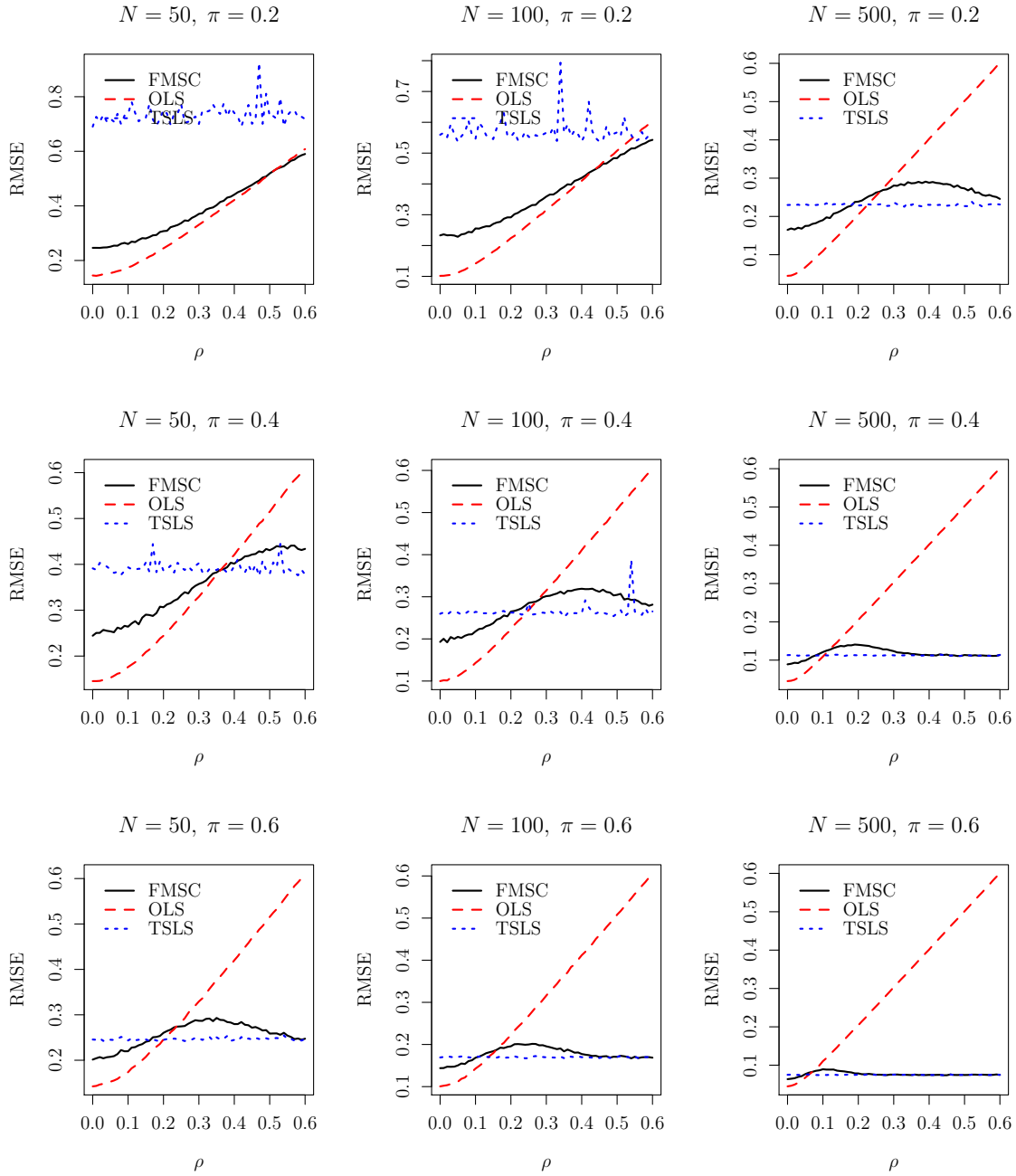


Figure 1: RMSE values for the two-stage least squares (TSLS) estimator, the ordinary least squares (OLS) estimator, and the post-Focused Moment Selection Criterion (FMSC) estimator based on 10,000 simulation draws from the DGP given in Equations 11–12 using the formulas described in Sections 2.4 and 3.3.

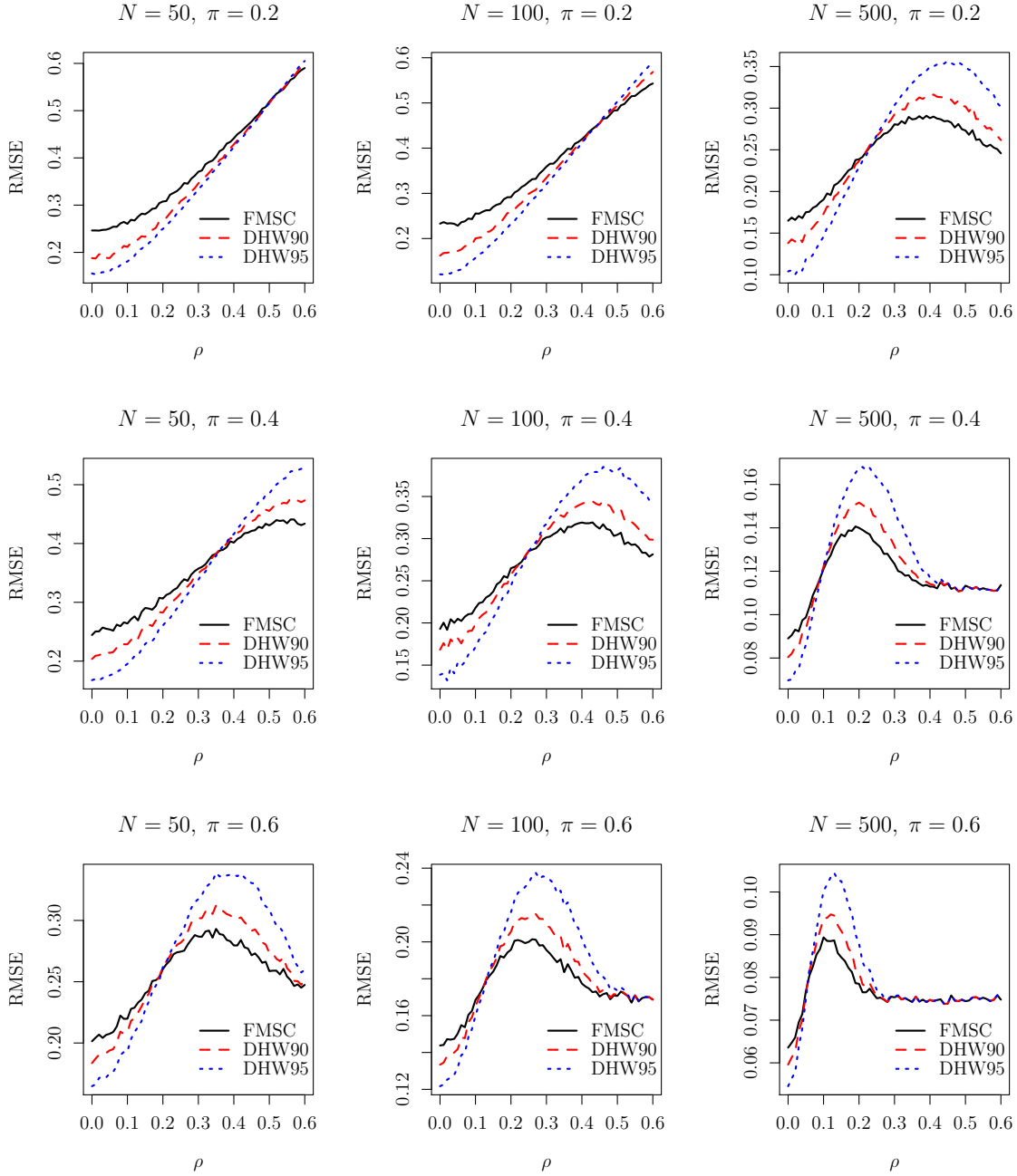


Figure 2: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator and Durbin-Hausman-Wu pre-test estimators with  $\alpha = 0.1$  (DHW90) and  $\alpha = 0.05$  (DHW95) based on 10,000 simulation draws from the DGP given in Equations 11–12 using the formulas described in Sections 2.4 and 3.3.

this section, let

$$\Xi_1 = [ I_p \quad 0_{p \times q} ], \quad \Xi_2 = [ 0_{q \times p} \quad I_q ] \quad (13)$$

where  $0_{m \times n}$  denotes an  $m \times n$  matrix of zeros and  $I_m$  denotes the  $m \times m$  identity matrix. Using this convention,  $Z_1 = Z\Xi_1'$  and  $Z_2 = Z\Xi_2'$ . In this example the valid estimator, defined in Assumption 2.4, is given by

$$\hat{\beta}_v = \left( X'Z_1\widetilde{W}_vZ_1'X \right)^{-1} X'Z_1\widetilde{W}_vZ_1'y \quad (14)$$

and we estimate  $\nabla_{\beta}\mu(\beta)$  with  $\nabla_{\beta}\mu(\hat{\beta}_v)$ . Similarly,

$$-\widehat{K}_S = n \left( X'Z\Xi_S'\widetilde{W}_S\Xi_SZ'X \right)^{-1} X'Z\Xi_S'\widetilde{W}_S$$

is the natural consistent estimator of  $-K_S$  in this setting.<sup>11</sup> Since  $\Xi_S$  is known, the only remaining quantities from Equation 9 are  $\widehat{\tau}$ ,  $\widehat{\Psi}$  and  $\widehat{\Omega}$ . The following result specializes Theorem 3.1 to the present example.

**Theorem 3.3.** *Let  $\widehat{\tau} = n^{-1/2}Z_2'(\mathbf{y} - X\widehat{\beta}_v)$  where  $\widehat{\beta}_v$  is as defined in Equation 14. Under Assumption 2.6 we have  $\widehat{\tau} \rightarrow_d \tau + \Psi M$  where  $M$  is defined in Theorem 2.4,*

$$\begin{aligned} \Psi &= [ -\Xi_2Q\Pi K_v \quad I_q ] \\ -K_v &= (\Pi'Q\Xi_1'W_v\Xi_1Q'\Pi)^{-1} \Pi'Q\Xi_1'W_v \end{aligned}$$

$W_v$  is the probability limit of the weighting matrix from Equation 14,  $I_q$  is the  $q \times q$  identity matrix,  $\Xi_1$  is defined in Equation 13, and  $Q$  is defined in Assumption 2.6.

Using the preceding result, I construct the asymptotically unbiased estimator  $\widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}'$  of  $\tau\tau'$  from

$$\widehat{\Psi} = \left[ -n^{-1}Z_2'X \left( -\widehat{K}_v \right) \quad I_q \right], \quad -\widehat{K}_v = n \left( X'Z_1\widetilde{W}_vZ_1'X \right)^{-1} X'Z_1\widetilde{W}_v$$

All that remains before substituting values into Equation 9 is to estimate  $\Omega$ . There are many possible ways to proceed, depending on the problem at hand and the assumptions one is willing to make. In the simulation and empirical examples discussed below I examine the TSLS estimator, that is  $\widetilde{W}_S = (\Xi_SZ'Z\Xi_S)^{-1}$ , and estimate  $\Omega$  as follows. For all specifications *except* the valid estimator  $\widehat{\beta}_v$ , I employ the centered, heteroskedasticity-consistent estimator

$$\widehat{\Omega}_S = \frac{1}{n} \sum_{i=1}^n u_i(\widehat{\beta}_S)^2 \mathbf{z}_{iS}\mathbf{z}_{iS}' - \left( \frac{1}{n} \sum_{i=1}^n u_i(\widehat{\beta}_S)\mathbf{z}_{iS} \right) \left( \frac{1}{n} \sum_{i=1}^n u_i(\widehat{\beta}_S)\mathbf{z}_{iS}' \right) \quad (15)$$

<sup>11</sup>The negative sign is squared in the FMSC expression and hence disappears. I write it here only to be consistent with the notation of Theorem 2.2.

where  $u_i(\beta) = y_i - \mathbf{x}'_i\beta$ ,  $\widehat{\beta}_S = (X'Z_S(Z'_SZ_S)^{-1}Z'_SX)^{-1}X'Z_S(Z'_SZ_S)^{-1}Z'_S\mathbf{y}$ ,  $\mathbf{z}_{iS} = \Xi_S\mathbf{z}_i$  and  $Z'_S = \Xi_SZ'$ . Centering allows moment functions to have non-zero means. While the local mis-specification framework implies that these means tend to zero in the limit, they are non-zero for any fixed sample size. Centering accounts for this fact, and thus provides added robustness. Since the valid estimator  $\widehat{\beta}_v$  has no asymptotic bias, the AMSE of any target parameter based on this estimator equals its asymptotic variance. Accordingly, I use

$$\widetilde{\Omega}_{11} = n^{-1} \sum_{i=1}^n u_i(\widehat{\beta}_v)^2 \mathbf{z}_{1i} \mathbf{z}'_{1i} \quad (16)$$

rather than the  $(p \times p)$  upper left sub-matrix of  $\widehat{\Omega}$  to estimate this quantity. This imposes the assumption that all instruments in  $Z_1$  are valid so that no centering is needed, providing greater precision.

### 3.6 Simulation Study: Choosing Instruments Example

I now evaluate the performance of FMSC in the Instrument Selection Example described in Section 3.5 using the following simulation design:

$$y_i = 0.5x_i + \epsilon_i \quad (17)$$

$$x_i = (z_{1i} + z_{2i} + z_{3i})/3 + \gamma w_i + v_i \quad (18)$$

for  $i = 1, 2, \dots, N$  where  $(\epsilon_i, v_i, w_i, z_{i1}, z_{2i}, z_{3i})' \sim \text{iid } N(0, \mathcal{V})$  with

$$\mathcal{V} = \begin{bmatrix} 1 & (0.5 - \gamma\rho) & \rho & 0 & 0 & 0 \\ (0.5 - \gamma\rho) & (8/9 - \gamma^2) & 0 & 0 & 0 & 0 \\ \rho & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 \end{bmatrix} \quad (19)$$

This setup keeps the variance of  $x$  fixed at one and the endogeneity of  $x$ ,  $Cor(x, \epsilon)$ , fixed at 0.5 while allowing the relevance,  $\gamma = Cor(x, w)$ , and endogeneity,  $\rho = Cor(w, \epsilon)$ , of the instrument  $w$  to vary. The instruments  $z_1, z_2, z_3$  are valid and exogenous: they have first-stage coefficients of 1/3 and are uncorrelated with the second stage error  $\epsilon$ . The additional instrument  $w$  is only relevant if  $\gamma \neq 0$  and is only exogenous if  $\rho = 0$ . Since  $x$  has unit variance, the first-stage R-squared for this simulation design is simply  $1 - \sigma_v^2 = 1/9 + \gamma^2$ . Hence, when  $\gamma = 0$ , so that  $w$  is irrelevant, the first-stage R-squared is just over 0.11. Increasing  $\gamma$  increases the R-squared of the first-stage. When  $\gamma = 0$ , this simulation design is a special case of the DGP used in the Section 3.4.

As in Section 3.4, the goal of moment selection in this exercise is to estimate the effect of  $x$  on  $y$ , as before 0.5, with minimum MSE. In this case, however, the choice is between two TSLS estimators rather than OLS and TSLS: the *valid* estimator uses



only  $z_1, z_2$ , and  $z_3$  as instruments, while the *full* estimator uses  $z_1, z_2, z_3$ , and  $w$ . The inclusion of  $z_1, z_2$  and  $z_3$  in both moment sets means that the order of over-identification is two for the valid estimator and three for the full estimator. Because the moments of the TSLS estimator only exist up to the order of over-identification (Phillips, 1980), this ensures that the small-sample MSE is well-defined.<sup>12</sup> All estimators in this section are calculated via TSLS without a constant term using the expressions from Section 3.5 and 20,000 simulation replications.

Figure 3 presents RMSE values for the valid estimator, the full estimator, and the post-FMSC estimator for various combinations of  $\gamma$ ,  $\rho$ , and  $N$ . The results are broadly similar to those from the OLS versus TSLS example presented in Figure 1. For any combination  $(\gamma, N)$  there is a positive value of  $\rho$  below which the full estimator yields a lower RMSE than the full estimator. As the sample size increases, this cutoff becomes smaller; as  $\gamma$  increases, it becomes larger. As in the OLS versus TSLS example, the post-FMSC estimator represents a compromise between the two estimators over which the FMSC selects. Unlike in the previous example, however, when  $N$  is sufficiently small there is a range of values for  $\rho$  within which the FMSC yields a lower RMSE than *both* the valid and full estimators. This comes from the fact that the valid estimator is quite erratic for small sample sizes: even with 20,000 simulation replications it exhibits a noticeable degree of simulation error for  $N = 50$ . Such behavior is unsurprising given that its first stage is not especially strong, R-squared  $\approx 11\%$ , and it has only two moments. In contrast, the full estimator has three moments and a stronger first stage. As in the OLS versus TSLS example, the post-FMSC estimator does not uniformly outperform the valid estimator for all parameter values, although it does for smaller sample sizes. The FMSC never performs substantially worse than the valid estimator, however, and often performs dramatically better. Again, the gains are particularly large for small sample sizes.

I now compare the FMSC to a number of alternative moment selection procedures. The first is a downward  $J$ -test, an informal but fairly common procedure for moment selection in practice. This procedure simply uses the full estimator unless it is rejected by a  $J$ -test. For robustness, I calculate the  $J$ -test statistic using a centered covariance matrix estimator, as in the FMSC formulas from section 3.5. Table 4 compares the RMSE of the post-FMSC estimator to that of the downward  $J$ -test with  $\alpha = 0.1$  (J90), and  $\alpha = 0.05$  (J95). Unlike the FMSC, the downward  $J$ -test is very badly behaved for small sample sizes, particularly for the smaller values of  $\gamma$ . For larger sample sizes, the relative performance of the FMSC and the  $J$ -test is quite similar to what we saw in Figure 2 for the OLS versus TSLS example: the  $J$ -test performs best for the smallest values of  $\rho$ , the FMSC performs best for moderate values, and the two procedures perform similarly for large values. Andrews (1999) considers a more formal moment selection procedure based on criteria of the form  $MSC(S) = J_n(S) - h(|S|)\kappa_n$ , where  $J_n(S)$  is the  $J$ -test statistic under moment set  $S$ ,  $-h(|S|)\kappa_n$  is a “bonus term” that

---

<sup>12</sup>Alternatively, one could use fewer instruments for the valid estimator and compare the results using *trimmed* MSE, as in Hansen (2013).

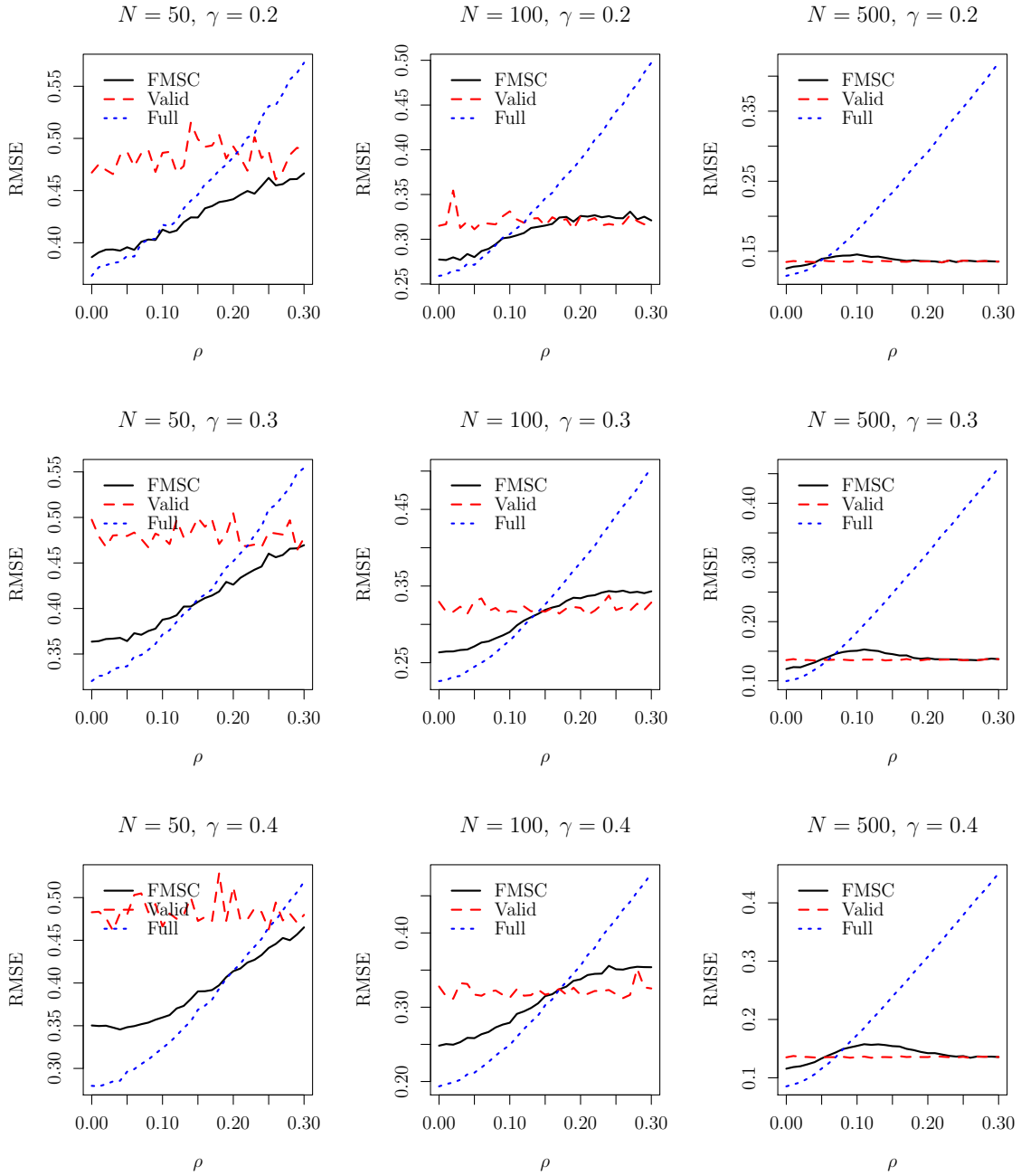


Figure 3: RMSE values for the valid estimator, including only  $(z_1, z_2, z_3)$ , the full estimator, including  $(z_1, z_2, z_3, w)$ , and the post-Focused Moment Selection Criterion (FMSC) estimator based on 20,000 simulation draws from the DGP given in Equations 18–19 using the formulas described in Sections 2.5 and 3.5.

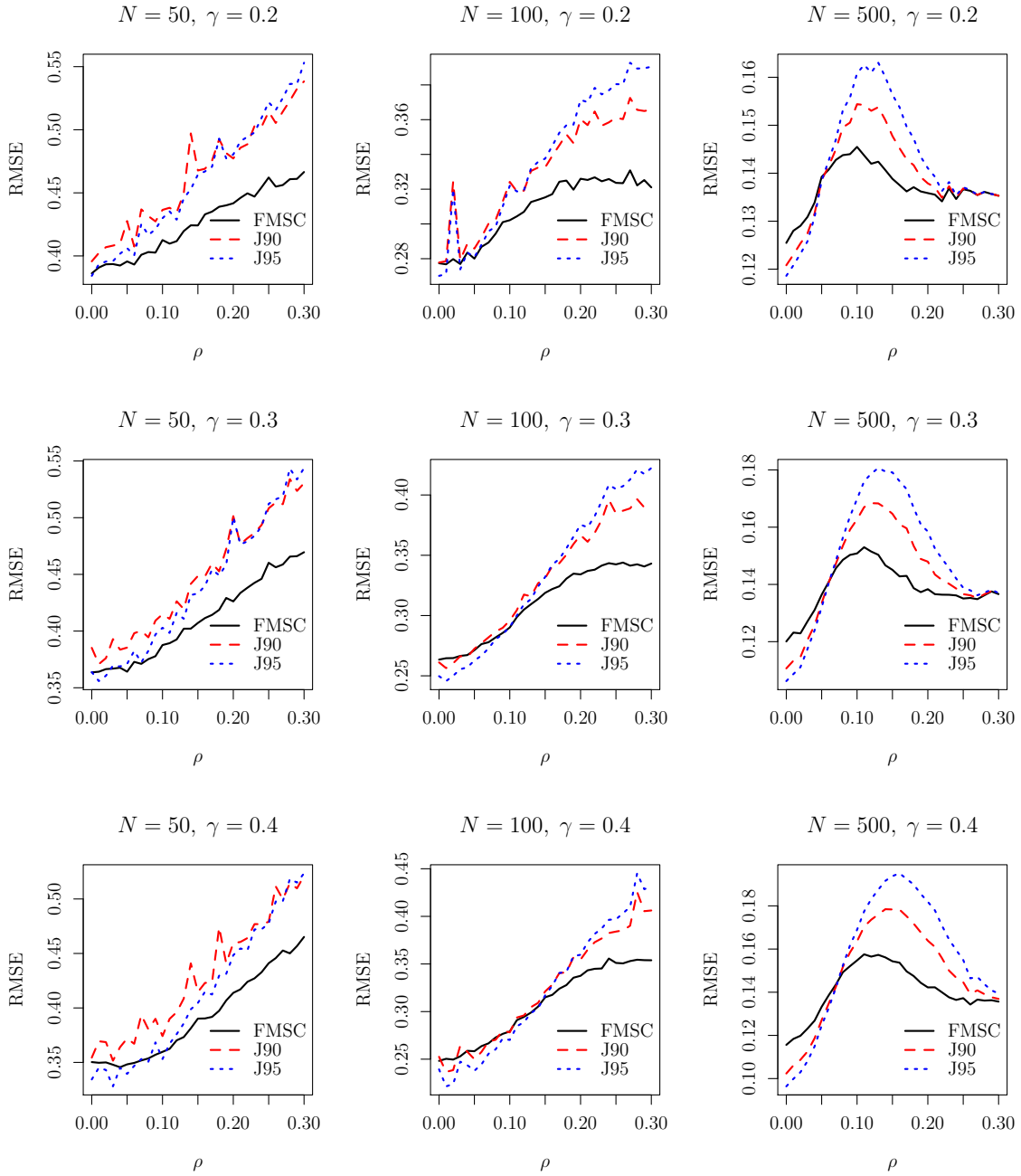


Figure 4: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator and the downward  $J$ -test estimator with  $\alpha = 0.1$  (J90) and  $\alpha = 0.05$  (J95) based on 20,000 simulation draws from the DGP given in Equations 18–19 using the formulas described in Sections 2.5 and 3.5.

rewards the inclusion of more moment conditions. For each member of this family we choose the moment set that *minimizes*  $MSC(S)$ . If we take  $h(|S|) = (p + |S| - r)$ , then  $\kappa_n = \log n$  gives a GMM analogue of Schwarz’s Bayesian Information Criterion (GMM-BIC) while  $\kappa_n = 2.01 \log \log n$  gives an analogue of the Hannan-Quinn Information Criterion (GMM-HQ), and  $\kappa_n = 2$  gives an analogue of Akaike’s Information Criterion (GMM-AIC). Like the maximum likelihood model selection criteria upon which they are based, the GMM-BIC and GMM-HQ are consistent provided that Assumption 2.3 holds, while the GMM-AIC, like the FMSC, is conservative.<sup>13</sup> Figure 5 gives the RMSE values for the post-FMSC estimator alongside those of the post-GMM-BIC, HQ and AIC estimators. As above, I calculate the  $J$ -test statistic using a centered covariance matrix estimator, following the recommendation of Andrews (1999). As described above, the  $J$ -test is quite erratic for small sample sizes. This problem carries over to the GMM-BIC, HQ and AIC: when  $N = 50$ , the FMSC has a uniformly smaller RMSE. As the sample size becomes larger, the classic tradeoff between consistent and conservative selection emerges. For the smallest values of  $\rho$  the consistent criteria outperform the conservative criteria; for moderate values the situation is reversed. The consistent criteria, however, have the highest worst-case RMSE. In this respect, GMM-BIC is the worst offender.

Because the Andrews-type criteria consider only instrument exogeneity, not relevance, Hall and Peixe (2003) suggest combining them with their canonical correlations information criterion (CCIC), which aims to detect and eliminate “redundant instruments.” Including such instruments, which add no information beyond that already contained in the other instruments, can lead to poor finite-sample performance in spite of the fact that the first-order limit distribution is unchanged. For the present simulation example, with a single endogenous regressor and no constant term, the CCIC takes the following simple form

$$CCIC(S) = n \log [1 - R_n^2(S)] + h(p + |S|)\kappa_n \quad (20)$$

where  $R_n^2(S)$  is the first-stage  $R^2$  based on instrument set  $S$  and  $h(p + |S|)\kappa_n$  is a penalty term (Jana, 2005). Instruments are chosen to *minimize* this criterion. When  $h(p + |S|) = (p + |S| - r)$ , setting  $\kappa_n = \log n$  gives the CCIC-BIC, while  $\kappa_n = 2.01 \log \log n$  gives the CCIC-HQ and  $\kappa_n = 2$  gives the CCIC-AIC. By combining the CCIC with an Andrews-type criterion, Hall and Peixe (2003) propose to first eliminate invalid instruments and then redundant ones. A combined GMM-BIC/CCIC-BIC criterion for the present example is to use the valid estimator unless both the GMM-BIC *and* CCIC-BIC select the full estimator. Combined HQ and AIC-type procedures can be defined analogously. Unfortunately, such combined criteria are of no use in the present simulation design: regardless of whether one uses the BIC, HQ or AIC version, the RMSE of the resulting is essentially identical to that of the valid estimator for all parameter values.<sup>14</sup>

<sup>13</sup>I discuss this further in Section 4 below.

<sup>14</sup>Details are available upon request.

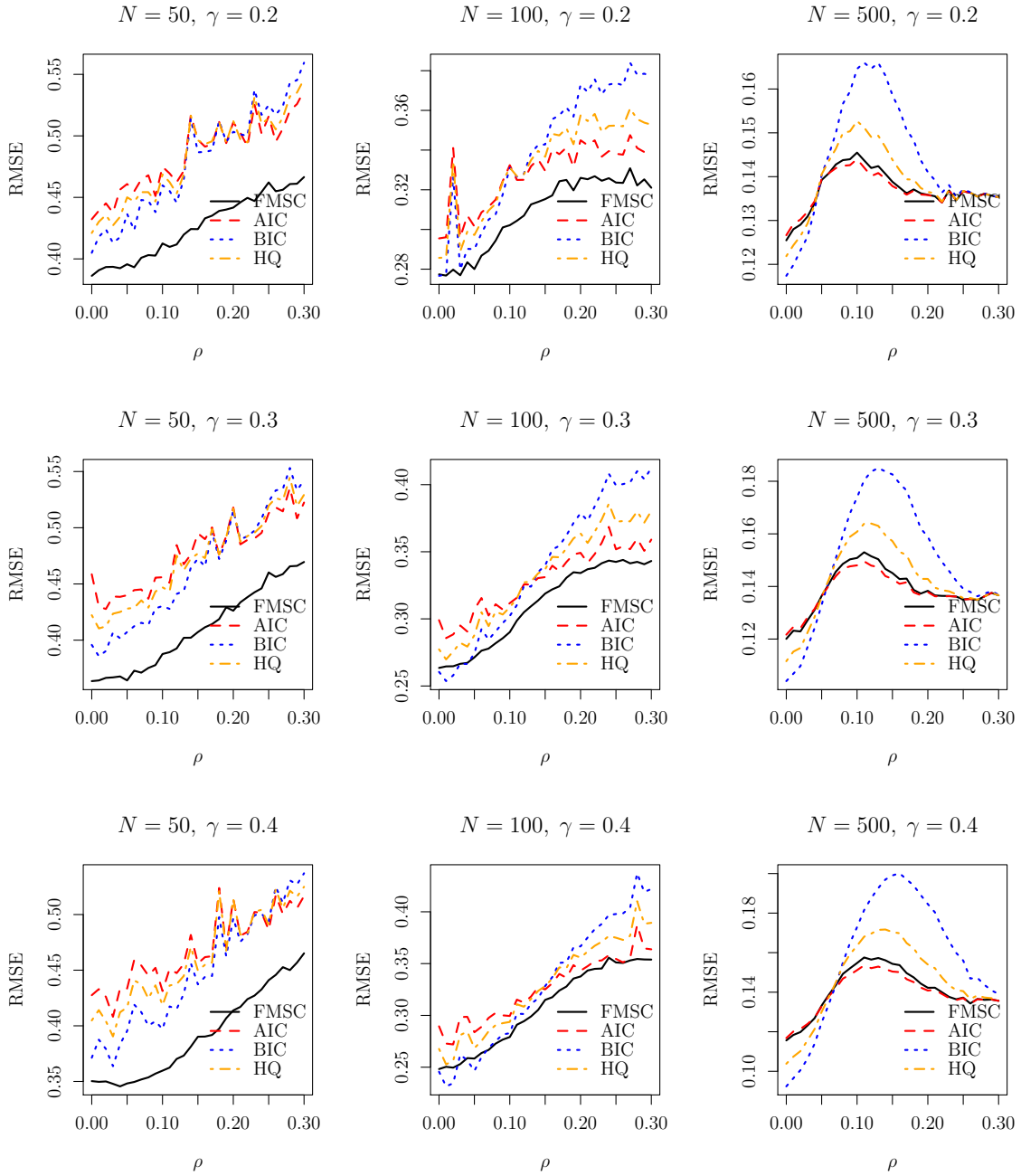


Figure 5: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator and the GMM-BIC, HQ, and AIC estimators based on 20,000 simulation draws from the DGP given in Equations 18-19 using the formulas described in Sections 2.5 and 3.5.

## 4 Moment Averaging & Post-Selection Estimators

The previous section derived the FMSC as an asymptotically unbiased estimator of the AMSE of a candidate estimator. Besides presenting simulation results, however, I have thus far said nothing about the sampling properties of the FMSC selection procedure itself. Because it is constructed from  $\hat{\tau}$ , the FMSC is a random variable, even in the limit. Combining Corollary 3.2 with Equation 9 gives the following.

**Corollary 4.1** (Limit Distribution of FMSC). *Under Assumptions 2.1, 2.2 and 2.4,  $FMSC_n(S) \rightarrow_d FMSC_S(\tau, M)$ , where*

$$FMSC_S(\tau, M) = \nabla_{\theta}\mu(\theta_0)'K_S\Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & B(\tau, M) \end{bmatrix} + \Omega \right\} \Xi_S'K_S'\nabla_{\theta}\mu(\theta_0)$$

$$B(\tau, M) = (\Psi M + \tau)(\Psi M + \tau)' - \Psi\Omega\Psi'.$$

This corollary implies that the FMSC is a “conservative” rather than “consistent” selection procedure. While this lack of consistency may sound like serious defect, it is in fact a desirable feature of the FMSC for two reasons. First, as discussed above, the goal of the FMSC is not to select only correctly specified moment conditions: it is to choose an estimator with a low finite-sample MSE as approximated by AMSE. In fact, the goal of consistent selection is very much at odds with that of controlling estimator risk. As explained by Yang (2005) and Leeb and Pötscher (2008), the worst-case risk of a consistent selection procedure *diverges* with sample size. This fact is readily apparent from the results of the simulation study from Section 3.6: the consistent criteria, GMM-BIC and HQ, have the highest worst-case RMSE, while the conservative criteria, FMSC and GMM-AIC, have the lowest. Second, while we know from both simulation studies (Demetrescu et al., 2011) and analytical examples (Leeb and Pötscher, 2005) that selection can dramatically change the sampling distribution of our estimators, invalidating traditional confidence intervals, the asymptotics of consistent selection give the misleading impression that this problem can be ignored. For example, Lemma 1.1 of Pötscher (1991, p. 168), which states that the limit distributions of an estimator pre- and post-consistent selection are identical, has been interpreted by some as evidence that consistent selection is innocuous. Pötscher (1991, pp. 179–180) makes it very clear, however, this result does not hold uniformly in the parameter space and hence “only creates an illusion of conducting valid inference” (Leeb and Pötscher, 2005, p. 22). While it is indeed true for *fixed* parameter values that consistent selection has a negligible effect on inference in large samples, small changes in parameter values can lead to large changes in coverage probabilities. In contrast to those of consistent selection, the asymptotics of conservative selection under local mis-specification provide a far more accurate picture of the distribution of post-selection estimators. The point is not that conservative criteria – such as the FMSC, GMM-AIC and  $J$ -test at a fixed significance level – are immune to the effects of selection on inference. Rather, it is that conservative criteria can be studied in an asymptotic framework that more accurately represents the effects encountered in finite samples.

So what exactly are these effects? The first is the addition of model selection uncertainty, a source of sampling variation that is not captured by traditional confidence intervals. If the data had been slightly different, we would have chosen a different set of moment conditions. Accordingly, any confidence interval that *conditions* on the selected moment conditions will be too short. The second effect comes from the fact that selection is carried out over *potentially mis-specified* estimators. Indeed, the FMSC explicitly sets out to use mis-specified moment conditions when doing so yields a favorable bias-variance tradeoff. Even if our goal were to eliminate all mis-specified moment conditions, for example by using a consistent criterion such as the GMM-BIC of [Andrews \(1999\)](#), in finite-samples we would not always be successful. Whenever there is a chance of selecting a mis-specified estimator, either on purpose or by mistake, a standard confidence interval that assumes all moment conditions to be correctly specified will be incorrectly centered.

To account for these two effects, we need a way to represent a *non-normal* sampling distribution in our limit theory, and this rules out consistent selection. The intuition is as follows. Each candidate estimator has a sampling distribution that is approximately normal by the central limit theorem. The post-selection estimator, however, is a data-dependent, randomly-weighted average of these. Hence, it follows a *mixture* distribution. Because they choose a single candidate with probability approaching one in the limit, consistent selection procedures make it impossible to represent this non-normality in the limit. In contrast, conservative selection procedures remain random even as the sample size goes to infinity, allowing us to derive a non-normal limit distribution and, ultimately, to carry out valid inference post-moment selection. To accomplish this, I first derive the asymptotic distribution of generic “moment average” estimators, extending the idea underlying the frequentist model average estimators of [Hjort and Claeskens \(2003\)](#). Moment average estimators are interesting in their own right and include, as a special case, a variety of post-conservative moment selection estimators including the FMSC. Because the resulting limit distributions are analytically intractable, I propose a two-step, simulation-based procedure for constructing confidence intervals with *uniform* asymptotic coverage guarantees.

## 4.1 Moment Average Estimators

A generic moment average estimator takes the form

$$\hat{\mu} = \sum_{S \in \mathcal{S}} \hat{\omega}_S \hat{\mu}_S \tag{21}$$

where  $\hat{\mu}_S = \mu(\hat{\theta}_S)$  is the estimator of the target parameter  $\mu$  under moment set  $S$ ,  $\mathcal{S}$  is the collection of all moment sets under consideration, and  $\hat{\omega}_S$  is shorthand for the value of a data-dependent weight function  $\hat{\omega}_S = \omega(\cdot, \cdot)$  evaluated at moment set  $S$  and the sample observations  $Z_{n1}, \dots, Z_{nn}$ . As above  $\mu(\cdot)$  is a  $\mathbb{R}$ -valued,  $Z$ -almost surely continuous function of  $\theta$  that is differentiable in an open neighborhood of  $\theta_0$ .

When  $\widehat{\omega}_S$  is an indicator, taking on the value one at the moment set moment set that minimizes some moment selection criterion,  $\widehat{\mu}$  is a post-moment selection estimator. To characterize the limit distribution of  $\widehat{\mu}$ , I impose the following conditions on  $\widehat{\omega}_S$ .

**Assumption 4.1** (Conditions on the Weights).

- (a)  $\sum_{S \in \mathcal{S}} \widehat{\omega}_S = 1$ , almost surely
- (b) For each  $S \in \mathcal{S}$ ,  $\widehat{\omega}_S \rightarrow_d \varphi_S(\tau, M)$ , an almost-surely continuous function of  $\tau$ ,  $M$  and consistently estimable constants only.

**Corollary 4.2** (Asymptotic Distribution of Moment-Average Estimators). *Under Assumption 4.1 and the conditions of Theorem 2.2,*

$$\sqrt{n}(\widehat{\mu} - \mu_0) \rightarrow_d \Lambda(\tau) = -\nabla_{\theta} \mu(\theta_0)' \left[ \sum_{S \in \mathcal{S}} \varphi_S(\tau, M) K_S \Xi_S \right] \left( M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right).$$

Notice that the limit random variable from Corollary 4.2, denoted  $\Lambda(\tau)$ , is a *randomly weighted average* of the multivariate normal vector  $M$ . Hence,  $\Lambda(\tau)$  is non-normal. This is precisely the behavior for which we set out to construct an asymptotic representation. The conditions of Assumption 4.1 are fairly mild. Requiring that the weights sum to one ensures that  $\widehat{\mu}$  is a consistent estimator of  $\mu_0$  and leads to a simpler expression for the limit distribution. While somewhat less transparent, the second condition is satisfied by weighting schemes based on a number of familiar moment selection criteria. It follows immediately from Corollary 4.1, for example, that the FMSC converges in distribution to a function of  $\tau$ ,  $M$  and consistently estimable constants only. The same is true for the  $J$ -test statistic, as seen from the following result.

**Theorem 4.1** (Distribution of  $J$ -Statistic under Local Mis-Specification). *Define the  $J$ -test statistic as per usual by  $J_n(S) = n \left[ \Xi_S f_n(\widehat{\theta}_S) \right]' \widehat{\Omega}^{-1} \left[ \Xi_S f_n(\widehat{\theta}_S) \right]$  where  $\widehat{\Omega}_S^{-1}$  is a consistent estimator of  $\Omega_S^{-1}$ . Then, under the conditions of Theorem 2.2, we have  $J_n(S) \rightarrow_d J_S(\tau, M)$  where*

$$J_S(\tau, M) = [\Omega_S^{-1/2}(M_S + \tau_S)]'(I - P_S)[\Omega_S^{-1/2}\Xi_S(M_S + \tau_S)],$$

$M_S = \Xi_S M$ ,  $\tau'_S = (0', \tau')\Xi'_S$ , and  $P_S$  is the projection matrix formed from the GMM identifying restrictions  $\Omega_S^{-1/2}F_S$ .

Hence, normalized weights constructed from almost-surely continuous functions of either the FMSC or the  $J$ -test statistic satisfy Assumption 4.1.

Post-selection estimators are merely a special cases of moment average estimators. To see why, consider the weight function

$$\widehat{\omega}_S^{MSC} = \mathbf{1} \left\{ \text{MSC}_n(S) = \min_{S' \in \mathcal{S}} \text{MSC}_n(S') \right\}$$



where  $MSC_n(S)$  is the value of some moment selection criterion evaluated at the sample observations  $Z_{n1} \dots, Z_{nn}$ . Now suppose  $MSC_n(S) \rightarrow_d MSC_S(\tau, M)$ , a function of  $\tau$ ,  $M$  and consistently estimable constants only. Then, so long as the probability of ties,  $P\{MSC_S(\tau, M) = MSC_{S'}(\tau, M)\}$ , is zero for all  $S \neq S'$ , the continuous mapping theorem gives

$$\widehat{\omega}_S^{MSC} \rightarrow_d \mathbf{1} \left\{ MSC_S(\tau, M) = \min_{S' \in \mathcal{S}} MSC_{S'}(\tau, M) \right\}$$

satisfying Assumption 4.1 (b). Thus, post-selection estimators based on the FMSC, the downward  $J$ -test procedure, GMM-BIC, GMM-HQ, and GMM-AIC all fall within the ambit of 4.2. GMM-BIC and GMM-HQ, however, are not particularly interesting under local mis-specification. Intuitively, because they aim to select all valid moment conditions w.p.a.1, we would expect that under Assumption 2.1 they simply choose the full moment set in the limit. The following result states that this intuition is correct.

**Theorem 4.2** (Consistent Criteria under Local Mis-Specification). *Consider a moment selection criterion of the form  $MSC(S) = J_n(S) - h(|S|)\kappa_n$ , where  $h$  is strictly increasing,  $\lim_{n \rightarrow \infty} \kappa_n = \infty$ , and  $\kappa_n = o(n)$ . Under the conditions of Theorem 2.2,  $MSC(S)$  selects the full moment set with probability approaching one.*

The preceding result is a special case of a more general phenomenon: consistent selection procedures cannot detect model violations of order  $O(n^{-1/2})$ . Because moment selection using the GMM-BIC or HQ leads to weights with a degenerate asymptotic distribution, one that does not capture the effects of selection on inference, I do not consider them further below.

## 4.2 Moment Averaging for the OLS versus TSLS Example

Although it is a special case of moment averaging, moment selection is a somewhat crude procedure: it gives full weight to the estimator that minimizes the moment selection criterion no matter how close its nearest competitor lies. Accordingly, when competing moment sets have similar criterion values in the population, sampling variation can be *magnified* in the selected estimator. Thus, it may be possible to achieve better performance by using smooth weights rather than discrete selection. In this section, I briefly explore this possibility.

The key question is: how should one construct the weights? One possibility is to adapt a proposal from [Buckland et al. \(1997\)](#), who suggest averaging a collection of competing maximum likelihood estimator with weights of the form

$$w_k = \exp(-I_k/2) / \sum_{i=1}^K \exp(-I_i/2)$$

where  $I_k$  is an information criterion evaluated for model  $k$ , and  $i$  indexes the set of  $K$  candidate models. This expression, constructed by an analogy with Bayesian model

averaging, gives more weight to models with lower values of the information criterion but non-zero weight to all models. Applying a slightly more general form of this idea, suggested by [Claeskens and Hjort \(2008b\)](#), to the moment selection criteria examined above, we might consider using weights of the form

$$\widehat{\omega}_S = \exp \left\{ -\frac{\kappa}{2} \text{MSC}(S) \right\} / \sum_{S' \in \mathcal{S}} \exp \left\{ -\frac{\kappa}{2} \text{MSC}(S') \right\}$$

where  $\text{MSC}(\cdot)$  is a moment selection criterion and the parameter  $\kappa \geq 0$  varies the uniformity of the weighting. As  $\kappa \rightarrow 0$  the weights become more uniform; as  $\kappa \rightarrow \infty$  they approach the moment selection procedure given by minimizing the corresponding criterion. Setting  $\kappa = 1$  gives the [Buckland et al. \(1997\)](#) weights.

Some preliminary simulation results, reported in an earlier draft of this paper, suggest that exponential weighting can indeed provide MSE improvements, so long as we choose an appropriate value for  $\kappa$ . And there's the rub: while it is easy to experiment with different values of  $\kappa$  in a simulation study, it is far from clear how to set about formulating a rule to choose  $\kappa$  in practice. In at least some applications, however, there is a compelling alternative to the exponential weighting scheme: one can instead derive weights *analytically* to minimize AMSE within the FMSC framework.<sup>15</sup> To illustrate this, I revisit the OLS versus TSLS example from Sections 2.4 and 3.3.

To begin, define an arbitrary weighted average of the OLS and TSLS estimators from Equations 3 and 4 by

$$\widetilde{\beta}(\omega) = \omega \widehat{\beta}_{OLS} + (1 - \omega) \widetilde{\beta}_{TSLS} \tag{22}$$

where  $\omega \in [0, 1]$  is the weight given to the OLS estimator. Since the weights sum to one, we have

$$\begin{aligned} \sqrt{n} \left[ \widehat{\beta}(\omega) - \beta \right] &= \begin{bmatrix} \omega & (1 - \omega) \end{bmatrix} \begin{bmatrix} \sqrt{n}(\widehat{\beta}_{OLS} - \beta) \\ \sqrt{n}(\widetilde{\beta}_{TSLS} - \beta) \end{bmatrix} \\ &\xrightarrow{d} N \left( \text{Bias} \left[ \widehat{\beta}(\omega) \right], \text{Var} \left[ \widehat{\beta}(\omega) \right] \right) \end{aligned}$$

by Theorem 2.3, where

$$\begin{aligned} \text{Bias} \left[ \widehat{\beta}(\omega) \right] &= \omega \left( \frac{\tau}{\sigma_x^2} \right) \\ \text{Var} \left[ \widehat{\beta}(\omega) \right] &= \frac{\sigma_\epsilon^2}{\sigma_x^2} \left[ (2\omega^2 - \omega) \left( \frac{\sigma_x^2}{\gamma^2} - 1 \right) + \frac{\sigma_x^2}{\gamma^2} \right] \end{aligned}$$

---

<sup>15</sup>I thank Bruce Hansen for suggesting this idea.

and accordingly

$$\text{AMSE} \left[ \widehat{\beta}(\omega) \right] = \omega^2 \left( \frac{\tau^2}{\sigma_x^4} \right) + (\omega^2 - 2\omega) \left( \frac{\sigma_\epsilon^2}{\sigma_x^2} \right) \left( \frac{\sigma_x^2}{\gamma^2} - 1 \right) + \frac{\sigma_\epsilon^2}{\gamma^2}. \quad (23)$$

The preceding expression is a globally convex function of  $\omega$ . Taking the first order condition and rearranging, we find that the unique global minimizer is

$$\omega^* = \arg \min_{\omega \in [0,1]} \text{AMSE} \left[ \widehat{\beta}(\omega) \right] = \left[ 1 + \frac{\tau^2/\sigma_x^4}{\sigma_\epsilon^2(1/\gamma^2 - 1/\sigma_x^2)} \right]^{-1} \quad (24)$$

In other words,

$$\omega^* = \left[ 1 + \frac{\text{ABIAS(OLS)}^2}{\text{AVAR(TSLS)} - \text{AVAR(OLS)}} \right]^{-1}$$

This expression has several important consequences. First, since the variance of the TSLS estimator is always strictly greater than that of the OLS estimator, the optimal value of  $\omega$  *cannot* be zero. No matter how strong the endogeneity of  $x$  as measured by  $\tau$ , we should always give some weight to the OLS estimator. Second, when  $\tau = 0$  the optimal value of  $\omega$  is one. If  $x$  is exogenous, OLS is strictly preferable to TSLS. Third, the optimal weights depend on the strength of the instruments  $\mathbf{z}$  as measured by  $\gamma$ . For a given value of  $\tau \neq 0$ , the stronger the instruments, the less weight we should give to OLS.

Equation 24 gives the AMSE-optimal weighted average of the OLS and TSLS estimators. To actually use the corresponding moment average estimator in practice, however, we need to estimate the unknowns. As discussed above in Section 3.3 the usual estimators of  $\sigma_x^2$  and  $\gamma$  remain consistent under local mis-specification, and the residuals from the TSLS estimator provide a robust estimator of  $\sigma_\epsilon^2$ . As before, the problem is estimating  $\tau^2$ . A natural idea is to substitute the asymptotically unbiased estimator that arises from Theorem 3.2, namely  $\widehat{\tau}^2 - \widehat{V}$ . The problem with this approach is that, while  $\tau^2$  is always greater than or equal to zero as is  $\widehat{\tau}^2$ , the difference  $\widehat{\tau}^2 - \widehat{V}$  could be negative, yielding a *negative* estimate of  $\text{ABIAS(OLS)}^2$ . To solve this problem, I borrow an idea from the literature on shrinkage estimation and use the *positive part* instead, namely  $\max\{0, \widehat{\tau}^2 - \widehat{V}\}$ , as in the positive-part James-Stein estimator. This ensures that the estimator of  $\omega^*$  lies inside the interval  $[0, 1]$ . Accordingly, I define

$$\widehat{\beta}_{AVG}^* = \widehat{\omega}^* \widehat{\beta}_{OLS} + (1 - \widehat{\omega}^*) \widetilde{\beta}_{TSLS} \quad (25)$$

where

$$\widehat{\omega}^* = \left[ 1 + \frac{\max\{0, (\widehat{\tau}^2 - \widehat{\sigma}_\epsilon^2 \widehat{\sigma}_x^2 (\widehat{\sigma}_x^2 / \widehat{\gamma}^2 - 1)) / \widehat{\sigma}_x^4\}}{\widehat{\sigma}_\epsilon^2 (1/\widehat{\gamma}^2 - 1/\widehat{\sigma}_x^2)} \right]^{-1} \quad (26)$$

Returning to the simulation experiment from Section 3.4, Figure 6 compares the finite-sample RMSE of the minimum-AMSE moment average estimator presented in Equations 25 and 26 to those of the FMSC and DHW pre-test estimators from Figure

2. The performance of the minimum-AMSE moment average estimator is strong: it provides the lowest worst-case RMSE and improves uniformly on the FMSC for all but the largest values of  $\rho$ . While a complete examination of minimum-AMSE moment averaging is beyond the scope of the present paper, these results suggest that it could prove extremely useful in practice. In the following section I show how one can construct a valid confidence interval for this and related procedures.

### 4.3 Valid Confidence Intervals

While Corollary 4.2 characterizes the limiting behavior of moment-average, and hence post-selection estimators, the limiting random variable  $\Lambda(\tau)$  is a complicated function of the normal random vector  $M$ . Because this distribution is analytically intractable, I adapt a suggestion from Claeskens and Hjort (2008b) and approximate it by simulation. The result is a conservative procedure that provides asymptotically valid confidence intervals for moment average and hence post-conservative selection estimators.<sup>16</sup>

First, suppose that  $K_S$ ,  $\varphi_S$ ,  $\theta_0$ ,  $\Omega$  and  $\tau$  were known. Then, by simulating from  $M$ , as defined in Theorem 2.2, the distribution of  $\Lambda(\tau)$ , defined in Corollary 4.2, could be approximated to arbitrary precision. To operationalize this procedure, substitute consistent estimators of  $K_S$ ,  $\theta_0$ , and  $\Omega$ , e.g. those used to calculate FMSC. To estimate  $\varphi_S$ , we first need to derive the limit distribution of  $\widehat{\omega}_S$ , the data-based weights specified by the user. As an example, consider the case of moment selection based on the FMSC. Here  $\widehat{\omega}_S$  is simply the indicator function

$$\widehat{\omega}_S = \mathbf{1} \left\{ \text{FMSC}_n(S) = \min_{S' \in \mathcal{S}} \text{FMSC}_n(S') \right\} \quad (27)$$

To estimate  $\varphi_S$ , first substitute consistent estimators of  $\Omega$ ,  $K_S$  and  $\theta_0$  into  $\text{FMSC}_S(\tau, M)$ , defined in Corollary 4.1, yielding,

$$\widehat{\text{FMSC}}_S(\tau, M) = \nabla_{\theta\mu}(\widehat{\theta})' \widehat{K}_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \widehat{\mathcal{B}}(\tau, M) \end{bmatrix} + \widehat{\Omega} \right\} \Xi_S' \widehat{K}_S' \nabla_{\theta\mu}(\widehat{\theta}) \quad (28)$$

where

$$\widehat{\mathcal{B}}(\tau, M) = (\widehat{\Psi}M + \tau)(\widehat{\Psi}M + \tau)' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi} \quad (29)$$

Combining this with Equation 27,

$$\widehat{\varphi}_S(\tau, M) = \mathbf{1} \left\{ \widehat{\text{FMSC}}_S(\tau, M) = \min_{S' \in \mathcal{S}} \widehat{\text{FMSC}}_{S'}(\tau, M) \right\}. \quad (30)$$

---

<sup>16</sup>Although I originally developed this procedure by analogy to Claeskens and Hjort (2008b), Leeb and Pötscher (2014) kindly pointed out that constructions of the kind given here have appeared elsewhere in the statistics literature, notably in Loh (1985), Berger and Boos (1994), and Silvapulle (1996). More recently, McCloskey (2012) uses a similar approach to study non-standard testing problems.

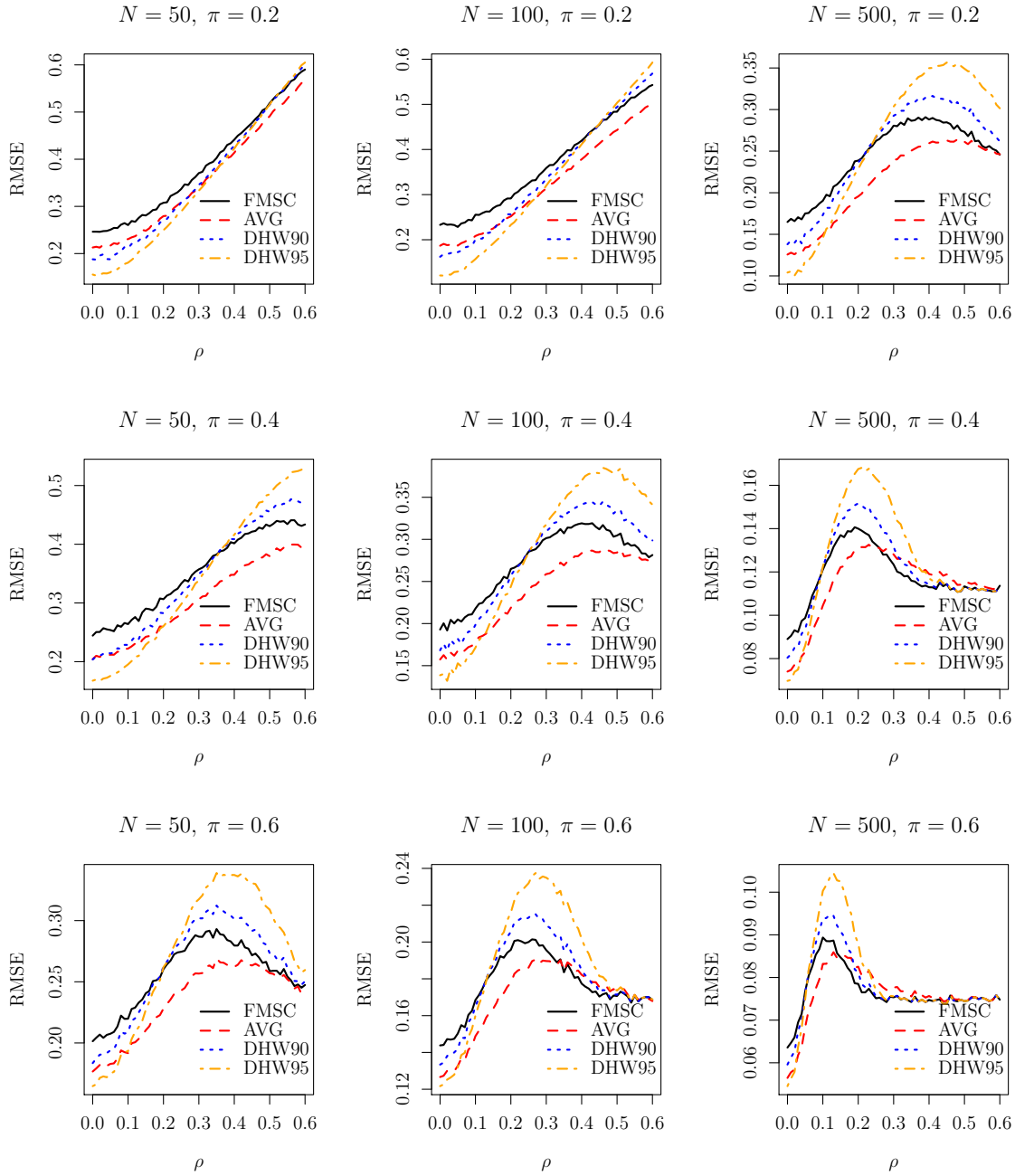


Figure 6: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator, Durbin-Hausman-Wu pre-test estimators with  $\alpha = 0.1$  (DHW90) and  $\alpha = 0.05$  (DHW95), and the minimum-AMSE averaging estimator, based on 10,000 simulation draws from the DGP given in Equations 11–12 using the formulas described in Sections 2.4, 3.3, and 4.2.

For GMM-AIC moment selection or selection based on a downward  $J$ -test,  $\varphi_S(\cdot, \cdot)$  may be estimated analogously, following Theorem 4.1.

Although simulating draws from  $M$ , defined in Theorem 2.2, requires only an estimate of  $\Omega$ , the limit  $\varphi_S$  of the weight function also depends on  $\tau$ . As discussed above, no consistent estimator of  $\tau$  is available under local mis-specification: the estimator  $\hat{\tau}$  has a non-degenerate limit distribution (see Theorem 3.1). Thus, substituting  $\hat{\tau}$  for  $\tau$  will give erroneous results by failing to account for the uncertainty that enters through  $\hat{\tau}$ . The solution is to use a two-stage procedure. First construct a  $100(1 - \delta)\%$  confidence region  $\mathcal{I}(\hat{\tau}, \delta)$  for  $\tau$  using Theorem 3.1. Then, for each  $\tau^* \in \mathcal{I}(\hat{\tau}, \delta)$  simulate from the distribution of  $\Lambda(\tau^*)$ , defined in Corollary 4.2, to obtain a *collection* of  $(1 - \alpha) \times 100\%$  confidence intervals indexed by  $\tau^*$ . Taking the lower and upper bounds of these yields a *conservative* confidence interval for  $\hat{\mu}$ , as defined in Equation 21. This interval has asymptotic coverage probability of *at least*  $(1 - \alpha - \delta) \times 100\%$ . The precise algorithm is as follows.

**Algorithm 4.1** (Simulation-based Confidence Interval for  $\hat{\mu}$ ).

1. For each  $\tau^* \in \mathcal{I}(\hat{\tau}, \delta)$ 
  - (i) Generate  $J$  independent draws  $M_j \sim N_{p+q}(0, \hat{\Omega})$
  - (ii) Set  $\Lambda_j(\tau^*) = -\nabla_{\theta} \mu(\hat{\theta})' \left[ \sum_{S \in \mathcal{S}} \hat{\varphi}_S(\tau^*, M_j) \hat{K}_S \Xi_S \right] (M_j + \tau^*)$
  - (iii) Using the draws  $\{\Lambda_j(\tau^*)\}_{j=1}^J$ , calculate  $\hat{a}(\tau^*)$ ,  $\hat{b}(\tau^*)$  such that

$$P \left\{ \hat{a}(\tau^*) \leq \Lambda(\tau^*) \leq \hat{b}(\tau^*) \right\} = 1 - \alpha$$

2. Set  $\hat{a}_{min}(\hat{\tau}) = \min_{\tau^* \in \mathcal{I}(\hat{\tau}, \delta)} \hat{a}(\tau^*)$  and  $\hat{b}_{max}(\hat{\tau}) = \max_{\tau^* \in \mathcal{I}(\hat{\tau}, \delta)} \hat{b}(\tau^*)$

3. The confidence interval for  $\mu$  is  $CI_{sim} = \left[ \hat{\mu} - \frac{\hat{b}_{max}(\hat{\tau})}{\sqrt{n}}, \hat{\mu} - \frac{\hat{a}_{min}(\hat{\tau})}{\sqrt{n}} \right]$

**Theorem 4.3** (Simulation-based Confidence Interval for  $\hat{\mu}$ ). *Let  $\hat{\Psi}$ ,  $\hat{\Omega}$ ,  $\hat{\theta}$ ,  $\hat{K}_S$ ,  $\hat{\varphi}_S$  be consistent estimators of  $\Psi$ ,  $\Omega$ ,  $\theta_0$ ,  $K_S$ ,  $\varphi_S$  and define*

$$\begin{aligned} \Delta_n(\hat{\tau}, \tau^*) &= (\hat{\tau} - \tau^*)' \left( \hat{\Psi} \hat{\Omega} \hat{\Psi}' \right)^{-1} (\hat{\tau} - \tau^*) \\ \mathcal{I}(\hat{\tau}, \delta) &= \left\{ \tau^* : \Delta_n(\hat{\tau}, \tau^*) \leq \chi_q^2(\delta) \right\} \end{aligned}$$

where  $\chi_q^2(\delta)$  denotes the  $1 - \delta$  quantile of a  $\chi^2$  distribution with  $q$  degrees of freedom. Then, the interval  $CI_{sim}$  defined in Algorithm 4.1 has asymptotic coverage probability no less than  $1 - (\alpha + \delta)$  as  $J, n \rightarrow \infty$ .

## 4.4 Simulation Study: Valid Confidence Intervals

To evaluate the the performance of Algorithm 4.1 in finite-samples, I now revisit the simulation experiments introduced above in Sections 3.4 and 3.6. All results in this section are based on 10,000 simulation replications from the appropriate DGP with  $\alpha = \delta = 0.05$ . For more computational details, see Appendix A. Coverage probabilities and relative widths are all given in percentage points, rounded to the nearest whole percent.

Tables 1b and 2b show the problem of ignoring moment selection by presenting the the actual coverage probability of a naïve 90%, post-FMSC confidence interval. The naïve procedure simply constructs a textbook 90% interval around the FMSC-selected estimator. As such it completely ignores both moment selection uncertainty, and the possibility that the selected estimator shows an asymptotic bias and is hence incorrectly centered. Table 1b reports results for the OLS versus TSLS simulation experiment, and Table 2b for the choosing instrumental variables simulation experiment. In each case, coverage probabilities can be made practically as close to zero as we like by choosing appropriate parameter values. This effect persists even for large sample sizes. At the same time, for many values in the parameter space the coverage probabilities are close or even equal to their nominal level. This is precisely the lack of uniformity described by Leeb and Pötscher (2005), and the reason why post-selection inference is such a knotty problem.

Table 3a gives the actual coverage probability of the conservative, 90% post-FMSC confidence interval, constructed according to Algorithm 4.1, for the OLS versus TSLS example. As seen from the table, these intervals achieve their nominal minimum coverage probability uniformly over the parameter space for all sample sizes. In general, however, they can be quite conservative, particularly for smaller values of  $\pi$ ,  $\rho$  and  $N$ : coverage probabilities never fall below 94% and occasionally exceed 99.5%.<sup>17</sup> Some conservatism is inevitable given the nature of the two-step procedure, which takes *worst-case* lower and upper bounds over a collection of intervals. In this example, however, the real culprit is the TSLS estimator itself. Table 1a presents coverage probabilities for a nominal 90% confidence interval for the TSLS estimator. Although this estimator is correctly specified and is not subject to moment selection uncertainty, so that the textbook asymptotic theory applies, coverage probabilities are far above their nominal level for smaller values of  $\pi$  even if  $N$  is fairly large, a manifestation of the weak instrument problem. This additional source of conservatism is inherited by the two-step post-FMSC intervals. Results for the minimum-AMSE moment average estimator, given in Table 3b, are broadly similar. Although the two-step intervals continue to achieve their nominal minimum coverage probability for all parameter values, they are quite conservative. If anything, the minimum-AMSE intervals are a bit more conservative than the post-FMSC intervals from Table 1a.

The worry, of course, is not conservatism as such but the attendant increase in

---

<sup>17</sup>Recall that coverage probabilities are given in percentage points, rounded to the nearest whole percent. Thus a value of 100, for example, in fact means  $\geq 99.5$ .

(a) Two-Stage Least Squares							(b) Naïve post-FMSC								
		$\rho$							$\rho$						
$N = 50$		0	0.1	0.2	0.3	0.4	0.5	$N = 50$		0	0.1	0.2	0.3	0.4	0.5
	0.1	98	98	96	93	89	82		0.1	88	80	58	30	11	4
	0.2	97	97	95	93	88	83		0.2	88	79	59	34	15	10
$\pi$	0.3	96	96	94	92	88	85	$\pi$	0.3	87	81	62	39	25	23
	0.4	94	93	93	91	89	87		0.4	86	80	66	46	38	43
	0.5	92	92	92	91	90	88		0.5	86	81	68	56	54	62
	0.6	91	91	90	90	90	88		0.6	85	81	72	66	67	75

		$\rho$							$\rho$						
$N = 100$		0	0.1	0.2	0.3	0.4	0.5	$N = 100$		0	0.1	0.2	0.3	0.4	0.5
	0.1	98	98	97	94	89	83		0.1	88	72	36	10	4	4
	0.2	96	96	95	92	89	85		0.2	87	74	40	17	13	19
$\pi$	0.3	94	94	93	91	89	87	$\pi$	0.3	86	74	45	29	32	45
	0.4	92	92	92	91	90	88		0.4	85	74	51	43	54	70
	0.5	91	91	90	90	89	89		0.5	85	76	59	57	70	84
	0.6	90	90	90	90	90	89		0.6	85	78	66	68	81	88

		$\rho$							$\rho$						
$N = 500$		0	0.1	0.2	0.3	0.4	0.5	$N = 500$		0	0.1	0.2	0.3	0.4	0.5
	0.1	96	96	94	93	90	86		0.1	87	31	8	12	16	24
	0.2	92	92	91	91	90	89		0.2	84	35	24	42	62	80
$\pi$	0.3	91	91	91	91	90	90	$\pi$	0.3	83	42	43	70	87	90
	0.4	90	90	91	90	90	90		0.4	84	49	62	86	90	90
	0.5	90	90	90	90	90	90		0.5	84	57	76	89	90	90
	0.6	90	91	90	90	90	90		0.6	86	66	84	90	90	90

Table 1: Coverage probabilities of nominal 90% CIs for the OLS versus TSLS simulation experiment from Section 3.4. All values are given in percentage points, rounded to the nearest whole percent, based on 10,000 simulation draws from the DGP given in Equations 11–12

confidence interval width. Accordingly, Tables 4a and 4b compare the median width of the simulation-based, two-step post-FMSC and minimum-AMSE intervals to that of the TSLS estimator. A value of 25, for example indicates that the simulation-based interval is 25% wider than the corresponding interval for the TSLS estimator. This comparison shows us the inferential cost of carrying out moment selection relative to simply using the correctly-specified TSLS estimator and calling it a day. We see immediately that moment selection is not a free lunch: the averaging and post-selection intervals are indeed wider than those of the TSLS estimator, sometimes considerably so. Intriguingly, although the minimum-AMSE intervals are generally more conservative than the post-FMSC intervals, they are also considerably shorter for nearly all parameter values.

Turning our attention now to the choosing instrumental variables simulation ex-



(a) Valid Estimator							(b) Naïve post-FMSC							
$N = 50$	$\rho$						$N = 50$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
	0.1	83	83	83	82	83	83	0.1	80	78	77	79	81	82
	0.2	83	83	83	83	84	83	0.2	79	75	73	74	79	81
$\gamma$	0.3	83	82	83	83	83	83	0.3	79	72	66	67	73	78
	0.4	82	83	84	83	83	84	0.4	78	71	62	59	66	75
	0.5	84	83	83	83	83	83	0.5	78	68	57	52	57	67
	0.6	83	83	83	83	82	82	0.6	77	68	54	47	50	60

$N = 100$	$\rho$						$N = 100$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
	0.1	86	87	87	87	86	86	0.1	84	82	84	86	86	86
	0.2	86	86	86	86	87	86	0.2	83	77	78	84	87	86
$\gamma$	0.3	86	86	87	87	87	87	0.3	82	74	71	79	85	86
	0.4	86	87	86	86	87	87	0.4	82	71	63	71	82	86
	0.5	87	87	86	86	86	86	0.5	81	69	56	62	76	84
	0.6	87	86	86	87	86	87	0.6	81	66	51	52	69	81

$N = 500$	$\rho$						$N = 500$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
	0.1	90	89	89	89	90	90	0.1	89	88	89	89	90	90
	0.2	89	90	90	90	90	90	0.2	87	84	90	90	90	90
$\gamma$	0.3	89	90	90	90	90	90	0.3	86	77	89	90	90	90
	0.4	89	90	90	89	90	90	0.4	85	67	88	89	90	90
	0.5	90	90	90	90	90	89	0.5	84	59	84	90	90	89
	0.6	89	89	89	90	90	90	0.6	83	52	77	89	90	90

Table 2: Coverage probabilities of nominal 90% CIs for the choosing instrumental variables simulation experiment described in Section 3.6. All values are given in percentage points, rounded to the nearest whole percent, based on 10,000 simulation draws from the DGP given in Equations 18–19.

periment from section 3.6, Table 5a gives the coverage probability and Table 5b the median relative width of the conservative, 90%, simulation-based, post-FMSC confidence interval. In this case, the width calculation is relative to the valid estimator, the TSLS estimator that includes the exogenous instruments  $z_1, z_2, z_3$  but excludes the potentially endogenous instrument  $w$ . In this example, the simulation-based intervals are far less conservative and, for the smallest sample sizes, sometimes fall slightly below their nominal minimum coverage probability. The worst case, 81% actual coverage compared to 90% nominal coverage, occurs when  $N = 50, \gamma = 0.6, \rho = 0.5$ . This problem stems from the fact that traditional interval for the valid estimator systematically under-covers when  $N = 50$  or 100, as we see from Table 2a. In spite of this, however, the simulation-based interval works well in this example. At the worst-case parameter values its median width is only 22% greater than that of the valid estimator.

(a) FMSC							(b) AMSE-Averaging Estimator								
$N = 50$	$\rho$						$N = 50$	$\rho$							
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5		
	0.1	100	100	99	99	98	97	0.1	100	100	100	99	98	96	
	0.2	99	99	99	99	98	97	0.2	100	100	100	99	98	96	
$\pi$	0.3	99	99	99	99	98	96	$\pi$	0.3	100	100	99	99	98	95
	0.4	98	98	98	98	98	95		0.4	99	99	99	98	97	94
	0.5	97	98	98	98	97	94		0.5	99	99	98	98	96	93
	0.6	97	97	97	97	96	94		0.6	98	98	98	97	96	93

$N = 100$	$\rho$						$N = 100$	$\rho$							
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5		
	0.1	100	99	99	99	99	98	0.1	100	100	100	100	99	97	
	0.2	99	99	99	99	99	97	0.2	100	100	100	99	98	95	
$\pi$	0.3	98	98	99	99	98	95	$\pi$	0.3	99	99	99	99	97	94
	0.4	97	97	98	98	97	94		0.4	99	99	99	98	96	94
	0.5	97	97	98	97	95	95		0.5	98	99	98	97	95	94
	0.6	97	97	97	96	95	96		0.6	98	98	97	96	95	95

$N = 500$	$\rho$						$N = 500$	$\rho$							
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5		
	0.1	99	99	99	99	99	96	0.1	100	100	100	99	98	95	
	0.2	97	97	98	99	97	94	0.2	99	99	99	98	96	94	
$\pi$	0.3	96	97	98	97	95	98	$\pi$	0.3	98	98	98	97	95	97
	0.4	96	97	97	95	98	98		0.4	98	98	97	96	96	97
	0.5	96	97	96	97	98	98		0.5	98	98	96	96	97	97
	0.6	96	97	95	97	97	96		0.6	98	97	96	96	97	96

Table 3: Coverage probabilities of simulation-based conservative 90% CIs for the OLS versus TSLS simulation experiment from Section 3.4. All values are given in percentage points, rounded to the nearest whole percent, based on 10,000 simulation draws from the DGP given in Equations 11–12.

On the whole, the simulation-based intervals constructed using Algorithm 4.1 work well, but there are two caveats. First, in situations where the standard first-order asymptotic theory begins to break down, such as a weak instruments example, the simulation-based intervals can inherit an under- or over-coverage problem from the valid estimator. It may be possible to address this problem by casting the moment selection problem explicitly within a weak identification framework. This could be a fruitful extension of the work presented here. Second, and more importantly, moment selection is not without cost: all of the simulation-based intervals are on average wider than a textbook confidence interval for the valid estimator. This raises an important question: why do moment selection at all? Wouldn't it be better to simply use the valid estimator and report a traditional confidence interval? This is a fair point. Moment selection is not a panacea and should not be employed mindlessly. The primary goal

of the the FMSC, however, is *estimation* rather than inference. And as shown above, using it can indeed provide a sizeable reduction in estimator RMSE relative to the valid estimator. Once the decision to carry out moment selection has been taken, however, one cannot simply ignore this fact and report the usual confidence intervals. Algorithm 4.1 provides a way to carry out honest inference post-selection and construct confidence intervals for non-standard estimators such as the minimum-AMSE averaging estimator from Section 4.2. The simulation results presented above show that it works fairly well in practice. Moreover, although formal moment selection is relatively rare in applied work, *informal* moment selection is extremely common. Downward  $J$ -tests, DHW tests and the like are a standard part of the applied econometrician’s toolkit. Because it can be employed to construct confidence intervals that account for the effect of such specification tests, Algorithm 4.1 can provide a valuable robustness check.

Although various impossibility results for post-selection inference, outlined in Leeb and Pötscher (2005), imply that conservative intervals are the best we can do if we desire inference with uniform coverage guarantees, the intervals presented above are the shortest possible. Note from Theorem 4.3 that *any* values of  $\alpha$  and  $\delta$  that sum to 0.1 yield an interval with asymptotic coverage probability no less than 90%. Accordingly, as suggested by Claeskens and Hjort (2008b) and investigated for a more general class of problems by McCloskey (2012), one could attempt to *optimize* confidence interval width by varying  $\alpha$  and  $\delta$ . The primary obstacle to this idea is computational. The simulation-based interval from Algorithm 4.1 is quite computationally intensive when the dimension of  $\tau$  is large. Adding another layer to the optimization procedure only makes things worse. Developing clever computational shortcuts to surmount this problem is an interesting topic for future research.

(a) post-FMSC Estimator

$N = 50$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	40	41	41	41	42	42
0.2	41	42	42	43	42	42
$\pi$ 0.3	42	43	43	43	43	43
0.4	43	43	43	43	43	43
0.5	43	42	42	42	42	41
0.6	41	41	40	40	39	38

(b) AMSE-Averaging Estimator

$N = 50$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	32	33	33	33	33	33
0.2	33	34	34	34	34	34
$\pi$ 0.3	34	35	34	34	35	34
0.4	35	35	35	35	34	34
0.5	36	36	35	34	34	33
0.6	36	35	35	33	32	32

$N = 100$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	40	41	41	41	41	42
0.2	42	42	42	42	43	44
$\pi$ 0.3	43	43	43	44	45	46
0.4	43	43	43	44	44	44
0.5	43	43	42	42	42	42
0.6	42	41	40	39	39	39

$N = 100$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	33	32	32	32	32	33
0.2	34	33	34	33	34	35
$\pi$ 0.3	35	35	34	35	35	35
0.4	35	35	35	35	35	35
0.5	36	36	35	34	33	33
0.6	36	35	34	32	32	32

$N = 500$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	40	41	42	42	43	46
0.2	42	43	45	47	49	51
$\pi$ 0.3	43	44	46	48	49	49
0.4	43	44	45	46	46	44
0.5	43	43	42	42	39	27
0.6	42	40	39	37	28	20

$N = 500$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	31	32	33	33	34	35
0.2	33	34	35	36	37	39
$\pi$ 0.3	35	35	36	36	37	38
0.4	35	35	35	36	36	36
0.5	36	35	34	34	34	31
0.6	36	33	32	32	30	25

Table 4: Median width of two-step, simulation-based conservative 90% CI relative to that of a traditional 90% CI for the TSLS estimator in the OLS versus TSLS example from section 3.4. All values are given in percentage points, rounded to the nearest whole percent, based on 10,000 simulation draws from the DGP given in Equations 11–12.

(a) Coverage Probability						
$N = 50$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	89	88	88	87	89	89
0.2	90	88	87	86	88	89
$\gamma$ 0.3	91	89	87	85	86	88
0.4	91	91	87	84	83	87
0.5	92	91	88	84	82	84
0.6	92	92	90	85	82	81

(b) Relative Median Width						
$N = 50$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	20	18	18	17	17	19
0.2	20	18	17	16	16	16
$\gamma$ 0.3	21	17	15	14	14	17
0.4	21	17	13	13	13	15
0.5	22	16	13	12	12	13
0.6	22	17	13	10	10	12

$N = 100$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	92	90	90	91	91	91
0.2	92	90	88	89	91	91
$\gamma$ 0.3	93	90	87	88	90	91
0.4	94	92	86	84	88	90
0.5	94	93	87	83	84	89
0.6	94	93	89	82	82	86

$N = 100$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	20	18	16	15	14	14
0.2	22	17	15	13	12	14
$\gamma$ 0.3	22	16	13	12	12	14
0.4	21	16	11	10	10	14
0.5	21	15	11	9	9	12
0.6	21	15	11	8	8	11

$N = 500$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	95	93	92	92	92	92
0.2	96	91	92	91	92	93
$\gamma$ 0.3	96	89	92	92	92	93
0.4	96	89	90	91	92	93
0.5	96	90	87	91	91	92
0.6	95	92	83	91	91	92

$N = 500$	$\rho$					
	0	0.1	0.2	0.3	0.4	0.5
0.1	22	17	13	10	8	7
0.2	22	15	10	7	6	9
$\gamma$ 0.3	21	13	8	6	7	12
0.4	20	13	7	6	7	11
0.5	19	13	8	6	6	9
0.6	20	13	8	5	6	8

Table 5: Performance of the simulation-based, conservative 90% post-FMSC confidence interval in the choosing instrumental variables simulation from Section 3.6. The left panel gives coverage probabilities, and the right panel gives median widths relative to that of a traditional 90% interval for the valid estimator. All values are given in percentage points, rounded to the nearest whole percent, based on 10,000 simulation draws from the DGP given in Equations 18–19.

## 5 Empirical Example: Geography or Institutions?

Carstensen and Gundlach (2006) address a controversial question from the development literature: does geography directly effect income after controlling for institutions? A number of well-known studies find little or no direct effect of geographic endowments (Acemoglu et al., 2001; Easterly and Levine, 2003; Rodrik et al., 2004). Sachs (2003), on the other hand, shows that malaria transmission, a variable largely driven by ecological conditions, directly influences the level of per capita income, even after controlling for institutions. Because malaria transmission is very likely endogenous, Sachs uses a measure of “malaria ecology,” constructed to be exogenous both to present economic conditions and public health interventions, as an instrument. Carstensen and Gundlach (2006) address the robustness of Sachs’s results using the following baseline regression for a sample of 44 countries:

$$\ln gdp_i = \beta_1 + \beta_2 \cdot institutions_i + \beta_3 \cdot malaria_i + \epsilon_i \quad (31)$$

Treating both institutions and malaria transmission as endogenous, they consider a variety of measures of each and a number of instrument sets. In each case, they find large negative effects of malaria transmission, lending support to Sach’s conclusion.

In this section, I revisit the instrument selection exercise given in Table 2 of Carstensen and Gundlach (2006) using the FMSC and corrected confidence intervals described above. I consider two questions. First, based on the FMSC methodology, which instruments should we choose if our target parameter is  $\beta_3$ , the effect of malaria transmission on per capita income? Does the selected instrument set change if our target parameter is  $\beta_2$ , the effect of institutions? Second, are the results robust to the effects of instrument selection on inference? All results are calculated by TSLS using the formulas from Section 3.5 and the variables described in Table 6, with  $\ln gdp$  as the outcome variable and  $rule$  and  $malfal$  as measures of institutions and malaria transmission.

To apply the FMSC to the present example, we require a minimum of two valid instruments besides the constant term. Based on the arguments given by Acemoglu et al. (2001) and Sachs (2003), I proceed under the assumption that  $\ln mort$  and  $maleco$ , measures of early settler mortality and malaria ecology, are exogenous. Rather than selecting over all 128 possible instrument sets, I consider eight specifications formed from the four instrument blocks defined by Carstensen and Gundlach (2006). The baseline block contains  $\ln mort$ ,  $maleco$  and a constant; the climate block contains  $frost$ ,  $humid$ , and  $latitude$ ; the Europe block contains  $eurfrac$  and  $engfrac$ ; and the openness block contains  $coast$  and  $trade$ . Full descriptions of these variables appear in Table 6. Table 7 lists the eight instrument sets considered here, along with TSLS estimates and traditional 95% confidence intervals for each.<sup>18</sup>

---

<sup>18</sup>The results presented in Table 7 are virtually identical to those of Carstensen and Gundlach (2006). The slight discrepancy in the results for the baseline instrument set, Panel 1, comes from the fact that Carstensen and Gundlach (2006) include an additional observation for this specification:

Name	Description	
<i>lngdpc</i>	Real GDP/capita at PPP, 1995 International Dollars	Outcome
<i>rule</i>	Institutional quality (Average Governance Indicator)	Regressor
<i>malfal</i>	Fraction of population at risk of malaria transmission, 1994	Regressor
<i>lnmort</i>	Log settler mortality (per 1000 settlers), early 19th century	Baseline
<i>maleco</i>	Index of stability of malaria transmission	Baseline
<i>frost</i>	Prop. of land receiving at least 5 days of frost in winter	Climate
<i>humid</i>	Highest temp. in month with highest avg. afternoon humidity	Climate
<i>latitude</i>	Distance from equator (absolute value of latitude in degrees)	Climate
<i>eurfrac</i>	Fraction of pop. that speaks major West. European Language	Europe
<i>engfrac</i>	Fraction of pop. that speaks English	Europe
<i>coast</i>	Proportion of land area within 100km of sea coast	Openness
<i>trade</i>	Log Frankel-Romer predicted trade share	Openness

Table 6: Description of variables for Empirical Example.

Table 8 presents FMSC and positive-part FMSC results for instrument sets 1–8 as defined in Table 7. The positive-part FMSC sets a negative squared asymptotic bias estimate to zero when estimating AMSE. When the squared bias estimate is positive, FMSC and positive-part FMSC coincide; when the squared bias estimate is negative, positive-part FMSC is strictly greater than FMSC. Additional simulation results for the choosing instrumental variables experiment from Section 3.6, available upon request, reveal that the positive-part FMSC never performs worse than the ordinary FMSC and sometimes performs better, suggesting that it should be preferred in real-world applications. For each criterion Table 8 presents two cases: the first takes the effect of *malfal*, a measure of malaria transmission, as the target parameter while the second uses the effect of *rule*, a measure of institutions. In each case the two best instrument sets are numbers 5 (baseline, climate and Europe) and 8 (all instruments).<sup>19</sup> When the target parameter is the coefficient on *malfal*, 8 is the clear winner under both the plain-vanilla and positive-part FMSC, leading to an estimate of  $-1.08$  for the effect of malaria transmission on per-capita income. When the target parameter is the coefficient on *rule*, however, instrument sets 5 and 8 are virtually indistinguishable. Indeed, while the plain-vanilla FMSC selects instrument set 8, leading to an estimate of 0.84 for the effect of institutions on per-capita income, the positive-part FMSC selects instrument set 5, leading to an estimate of 0.93. Thus the FMSC methodology shows that, while helpful for estimating the effect of malaria transmission, the openness instruments *coast* and *trade* provide essentially no additional information for studying the effect of institutions.

---

Vietnam. Because the Climate, Europe and Openness instrument blocks are missing for this country, I exclude it in the interest of consistency.

<sup>19</sup>In Table 2 of their paper, Carstensen and Gundlach (2006) report results for selection over instrument sets 2–4 and 8 using the GMM-BIC and HQ of Andrews (1999). Both favor instrument set 8. They do not consider instrument sets 5–7.

	1		2		3		4	
	<i>rule</i>	<i>malfal</i>	<i>rule</i>	<i>malfal</i>	<i>rule</i>	<i>malfal</i>	<i>rule</i>	<i>malfal</i>
coeff.	0.89	-1.04	0.97	-0.90	0.81	-1.09	0.86	-1.14
SE	0.18	0.31	0.16	0.29	0.16	0.29	0.16	0.27
lower	0.53	-1.66	0.65	-1.48	0.49	-1.67	0.55	-1.69
upper	1.25	-0.42	1.30	-0.32	1.13	-0.51	1.18	-0.59
	Baseline		Baseline Climate		Baseline		Baseline	
					Openness			
							Europe	

	5		6		7		8	
	<i>rule</i>	<i>malfal</i>	<i>rule</i>	<i>malfal</i>	<i>rule</i>	<i>malfal</i>	<i>rule</i>	<i>malfal</i>
coeff.	0.93	-1.02	0.86	-0.98	0.81	-1.16	0.84	-1.08
SE	0.15	0.26	0.14	0.27	0.15	0.27	0.13	0.25
lower	0.63	-1.54	0.59	-1.53	0.51	-1.70	0.57	-1.58
upper	1.22	-0.49	1.14	-0.43	1.11	-0.62	1.10	-0.58
	Baseline		Baseline		Baseline		Baseline	
	Climate		Climate				Climate	
			Openness		Openness		Openness	
	Europe				Europe		Europe	

Table 7: Two-stage least squares estimation results for all instrument sets.

Table 9 presents three alternative post-selection confidence intervals for each of the instrument selection exercises from Table 8: Naïve, 1-Step, and 2-Step. The Naïve intervals are standard, nominal 95% confidence intervals that ignore the effects of instrument selection. These are constructed by identifying the selected instrument set from Table 8 and simply reporting the corresponding nominal 95% interval from Table 7 unaltered. The 1-Step intervals are simulation-based nominal 95% intervals constructed using a simplified, and less computationally intensive, version of the procedure given in Algorithm 4.1. Rather than taking the minimum lower confidence limit and maximum upper confidence limit over all values in a given confidence region for  $\tau$ , this procedure simply assumes that the estimated value  $\hat{\tau}$  is exactly correct, and generates simulations for  $\Lambda$  under this assumption. Neither the Naïve nor the 1-Step procedures yield valid 95% confidence intervals. They are provided merely for comparison with the 2-Step procedure, which fully implements Algorithm 4.1 with  $\alpha = \delta = 0.05$  and  $J = 10,000$ . As explained above, the 2-Step interval is guaranteed to have asymptotic coverage probability of at least  $1 - \alpha - \delta$ , in this case 90%. From the 2-Step intervals, we see that the results of Carstensen and Gundlach (2006) are extremely robust. There is no evidence that accounting for the effects of instrument selection changes our conclusions



	$\mu = malfal$			$\mu = rule$		
	FMSC	posFMSC	$\hat{\mu}$	FMSC	posFMSC	$\hat{\mu}$
(1) Valid	3.03	3.03	-1.04	1.27	1.27	0.89
(2) Climate	3.07	3.07	-0.90	1.00	1.00	0.97
(3) Openness	2.30	2.42	-1.09	1.21	1.21	0.81
(4) Europe	1.82	2.15	-1.14	0.52	0.73	0.86
(5) Climate, Europe	0.85	2.03	-1.02	0.25	0.59	0.93
(6) Climate, Openness	1.85	2.30	-0.98	0.45	0.84	0.86
(7) Openness, Europe	1.63	1.80	-1.16	0.75	0.75	0.81
(8) Full	0.53	1.69	-1.08	0.23	0.62	0.84

Table 8: FMSC and and positive-part FMSC values corresponding to the instrument sets from Table 7

about the sign or significance of *malfal* or *rule*.

	$\mu = malfal$		$\mu = rule$	
	FMSC	posFMSC	FMSC	posFMSC
Naïve	(-1.58, -0.58)	(-1.58, -0.58)	(0.57, 1.10)	(0.63, 1.22)
1-Step	(-1.52, -0.67)	(-1.51, -0.68)	(0.57, 1.08)	(0.68, 1.17)
2-Step	(-1.62, -0.55)	(-1.62, -0.55)	(0.49, 1.18)	(0.58, 1.27)

Table 9: Post-selection CIs for the instrument selection exercise from Table 8.

Given the simplicity of the second-stage regression and the relatively small sample size, one might reasonably wonder whether this empirical example represents a realistic proof of concept for FMSC instrument selection and post-selection inference. The answer is yes: this example was selected precisely because it is as simple as possible in those respects that matter least for the FMSC and confidence interval calculations yet fully realistic in those respects that matter most. The computational complexity of the procedures described above is determined almost entirely by the dimension,  $q$ , of  $\tau$ , in this case the number of suspect instruments in the full instrument set. This is because the 2-Step confidence interval procedure requires us to carry out two  $q$ -dimensional constrained optimization problems with a stochastic objective function: one for the lower confidence limit and another for the upper confidence limit. Next most important is  $J$ , the number of simulation draws used, as this quantity determines the computational cost of a single evaluation of the objective function. In contrast, the number of instrument sets under consideration is considerably less important because we can pre-compute all associated quantities: they do not vary with  $\tau^*$ . Least important of all are the number of first-stage regressors and the sample size. For all practical purposes these are irrelevant. To put this another way, carrying out FMSC or positive-part FMSC selection is trivial; constructing valid confidence intervals, however, is challenging. With seven suspect instruments, 10000 simulation draws, and

eight instrument sets under consideration, this example presents the kind of computational challenge that one would reasonably expect to encounter in practice. Using a more complicated second-stage, more observations, or even double as many instrument sets would make practically no difference. In spite of the difficulties involved, problems of this magnitude are well within the ability of a standard desktop computer using off-the-shelf optimization routines. Running on a single core, it took just under eleven minutes to generate all of the empirical results in this paper. Note that this involves constructing *four* 2-Step confidence intervals. For more computational details, including a description of the packages used, see Appendix A.

## 6 Conclusion

This paper has introduced the FMSC, a proposal to choose moment conditions using AMSE. The criterion performs well in simulations, and the framework used to derive it allows us to construct valid confidence intervals for post-selection and moment-average estimators. Although computationally intensive, this procedure remains feasible for problems of a realistic scale without the need for specialized computing resources, as demonstrated in the empirical example above. Moment selection is not a panacea, but the FMSC and related confidence interval procedures can yield sizeable benefits in empirically relevant settings, making them a valuable complement to existing methods. While the discussion here concentrates on two cross-section examples, the FMSC could prove useful in any context in which moment conditions arise from more than one source. In a panel model, for example, the assumption of contemporaneously exogenous instruments may be plausible while that of predetermined instruments is more dubious. Using the FMSC, we could assess whether the extra information contained in the lagged instruments outweighs their potential invalidity. Work in progress explores this idea in both static and dynamic panel settings by extending the FMSC to allow for simultaneous moment and model selection. Other potentially fruitful extensions include the consideration of risk functions other than MSE, and an explicit treatment of weak identification and many moment conditions.

## A Computational Details

This paper is fully replicable using freely available, open-source software. For full source code and replication details, see <https://github.com/fditraglia/fmsc>. Results for the simulation studies and empirical example were generated using R version 3.1.0 (R Core Team, 2014) and C++ via the Rcpp (Eddelbuettel, 2013; Eddelbuettel and François, 2011) and RcppArmadillo (Eddelbuettel and Sanderson, 2014) packages, versions 0.11.2 and 0.4.300.8.0, respectively. RcppArmadillo version 0.4.300.8.0 provides an interface to version 4.300 of the Armadillo C++ linear algebra library (Sanderson, 2010). All figures in the paper were converted to tikz using version 0.7.0 of the

tikzDevice package (Sharpsteen and Bracken, 2013). The simulation-based confidence intervals from Section 4.4 were calculated using Algorithm 4.1 with  $J = 1000$  by searching over a grid of 100 equally spaced values within a 95% confidence interval for the scalar  $\tau$ . In contrast, the simulation-based intervals for the empirical example from Section 5 were constructed with  $J = 10,000$  using a mesh-adaptive search algorithm provided by version 3.6 of the NOMAD C++ optimization package (Abramson et al., 2013; Audet et al., 2009; Le Digabel, 2011), called from R using version 0.15-22 of the crs package (Racine and Nie, 2014). TSLs results for Table 7 were generated using version 3.1-4 of the sem package (Fox et al., 2014).

## B Proofs

**Proof of Theorems 2.1, 2.2.** Essentially identical to the proofs of Newey and McFadden (1994) Theorems 2.6 and 3.1.  $\square$

**Proof of Theorems 2.3, 2.4.** The proofs of both results are similar and standard, so we provide only a sketch of the argument for Theorem 2.4. First substitute the DGP into the expression for  $\hat{\beta}_S$  and rearrange so that the left-hand side becomes  $\sqrt{n}(\beta_S - \beta)$ . The right-hand side has two factors: the first converges in probability to  $-K_S$  by an  $L_2$  argument and the second converges in distribution to  $M + (0', \tau)'$  by the Lindeberg-Feller Central Limit Theorem.  $\square$

**Proof of Theorem 3.1.** By a mean-value expansion:

$$\begin{aligned}\hat{\tau} &= \sqrt{n}h_n(\hat{\theta}_v) = \sqrt{n}h_n(\theta_0) + H\sqrt{n}(\hat{\theta}_v - \theta_0) + o_p(1) \\ &= -HK_v\sqrt{n}g_n(\theta_0) + \mathbf{I}_q\sqrt{n}h_n(\theta_0) + o_p(1) \\ &= \Psi\sqrt{n}f_n(\theta_0) + o_p(1)\end{aligned}$$

The result follows since  $\sqrt{n}f_n(\theta_0) \rightarrow_d M + (0', \tau)'$  under Assumption 2.2 (h).  $\square$

**Proof of Corollary 3.2.** By Theorem 3.1 and the Continuous Mapping Theorem, we have  $\hat{\tau}\hat{\tau}' \rightarrow_d UU'$  where  $U = \Psi M + \tau$ . Since  $E[M] = 0$ ,  $E[UU'] = \Psi\Omega\Psi' + \tau\tau'$ .  $\square$

**Proof of Corollary 4.2.** Because the weights sum to one

$$\sqrt{n}(\hat{\mu} - \mu_0) = \sqrt{n}\left[\left(\sum_{S \in \mathcal{S}} \hat{\omega}_S \hat{\mu}_S\right) - \mu_0\right] = \sum_{S \in \mathcal{S}} [\hat{\omega}_S \sqrt{n}(\hat{\mu}_S - \mu_0)].$$

By Corollary 3.1, we have

$$\sqrt{n}(\hat{\mu}_S - \mu_0) \rightarrow_d -\nabla_{\theta}\mu(\theta_0)'K_S\Xi_S\left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix}\right)$$

and by the assumptions of this Corollary we find that  $\widehat{\omega}_S \rightarrow_d \varphi_S(\tau, M)$  for each  $S \in \mathcal{S}$ , where  $\varphi_S(\tau, M)$  is a function of  $M$  and constants only. Hence  $\widehat{\omega}_S$  and  $\sqrt{n}(\widehat{\mu}_S - \mu_0)$  converge jointly in distribution to their respective functions of  $M$ , for all  $S \in \mathcal{S}$ . The result follows by application of the Continuous Mapping Theorem.  $\square$

**Proof of Theorem 4.1.** By a mean-value expansion,

$$\sqrt{n} \left[ \Xi_S f_n \left( \widehat{\theta}_S \right) \right] = \sqrt{n} \left[ \Xi_S f_n \left( \theta_0 \right) \right] + F_S \sqrt{n} \left( \widehat{\theta}_S - \theta_0 \right) + o_p(1).$$

Since  $\sqrt{n} \left( \widehat{\theta}_S - \theta_0 \right) \rightarrow_p - \left( F'_S W_S F_S \right)^{-1} F'_S W_S \sqrt{n} \left[ \Xi_S f_n \left( \theta_0 \right) \right]$ , we have

$$\sqrt{n} \left[ \Xi_S f_n \left( \widehat{\theta}_S \right) \right] = \left[ I - F_S \left( F'_S W_S F_S \right)^{-1} F'_S W_S \right] \sqrt{n} \left[ \Xi_S f_n \left( \theta_0 \right) \right] + o_p(1).$$

Thus, for estimation using the efficient weighting matrix

$$\widehat{\Omega}_S^{-1/2} \sqrt{n} \left[ \Xi_S f_n \left( \widehat{\theta}_S \right) \right] \rightarrow_d \left[ I - P_S \right] \Omega_S^{-1/2} \Xi_S \left( M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

by Assumption 2.2 (h), where  $\widehat{\Omega}_S^{-1/2}$  is a consistent estimator of  $\Omega_S^{-1/2}$  and  $P_S$  is the projection matrix based on  $\Omega_S^{-1/2} F_S$ , the identifying restrictions.<sup>20</sup> The result follows by combining and rearranging these expressions.  $\square$

**Proof of Theorem 4.2.** Let  $S_1$  and  $S_2$  be arbitrary moment sets in  $\mathcal{S}$  and let  $|S|$  denote the cardinality of  $S$ . Further, define  $\Delta_n(S_1, S_2) = MSC(S_1) - MSC(S_2)$ . By Theorem 4.1,  $J_n(S) = O_p(1)$ ,  $S \in \mathcal{S}$ , thus

$$\begin{aligned} \Delta_n(S_1, S_2) &= [J_n(S_1) - J_n(S_2)] - [h(p + |S_1|) - h(p + |S_2|)] \kappa_n \\ &= O_p(1) - C \kappa_n \end{aligned}$$

where  $C = [h(p + |S_1|) - h(p + |S_2|)]$ . Since  $h$  is strictly increasing,  $C$  is positive for  $|S_1| > |S_2|$ , negative for  $|S_1| < |S_2|$ , and zero for  $|S_1| = |S_2|$ . Hence:

$$\begin{aligned} |S_1| > |S_2| &\implies \Delta_n(S_1, S_2) \rightarrow -\infty \\ |S_1| = |S_2| &\implies \Delta_n(S_1, S_2) = O_p(1) \\ |S_1| < |S_2| &\implies \Delta_n(S_1, S_2) \rightarrow \infty \end{aligned}$$

The result follows because the full moment set contains more moment conditions than any other moment set  $S$ .  $\square$

**Proof of Theorem 4.3.** By Theorem 3.1 and Corollary 4.2,

$$P \{ \mu_0 \in CI_{sim} \} \rightarrow P \{ a_{min} \leq \Lambda(\tau) \leq b_{max} \}$$

---

<sup>20</sup>See Hall (2005), Chapter 3.

where  $a(\tau^*), b(\tau^*)$  define a collection of  $(1 - \alpha) \times 100\%$  intervals indexed by  $\tau^*$ , each of which is constructed under the assumption that  $\tau = \tau^*$

$$P \{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} = 1 - \alpha$$

and we define the shorthand  $a_{min}, b_{max}$  as follows

$$\begin{aligned} a_{min}(\Psi M + \tau) &= \min \{a(\tau^*): \tau^* \in \mathcal{T}(\Psi M + \tau, \delta)\} \\ b_{max}(\Psi M + \tau) &= \max \{b(\tau^*): \tau^* \in \mathcal{T}(\Psi M + \tau, \delta)\} \\ \mathcal{T}(\Psi M + \tau, \delta) &= \{\tau^*: \Delta(\tau, \tau^*) \leq \chi_q^2(\delta)\} \\ \Delta(\tau, \tau^*) &= (\Psi M + \tau - \tau^*)'(\Psi \Omega \Psi')^{-1}(\Psi M + \tau - \tau^*) \end{aligned}$$

Now, let  $A = \{\Delta(\tau, \tau) \leq \chi_q^2(\delta)\}$  where  $\chi_q^2(\delta)$  is the  $1 - \delta$  quantile of a  $\chi_q^2$  random variable. This is the event that the *limiting version* of the confidence region for  $\tau$  contains the true bias parameter. Since  $\Delta(\tau, \tau) \sim \chi_q^2$ ,  $P(A) = 1 - \delta$ . For every  $\tau^* \in \mathcal{T}(\Psi M + \tau, \delta)$  we have

$$P[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A] + P[\{a(\tau^*) \leq \Lambda(\tau) \leq b(\tau^*)\} \cap A^c] = 1 - \alpha$$

by decomposing  $P \{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\}$  into the sum of mutually exclusive events. But since

$$P[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A^c] \leq P(A^c) = \delta$$

we see that

$$P[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A] \geq 1 - \alpha - \delta$$

for every  $\tau^* \in \mathcal{T}(\Psi M + \tau, \delta)$ . Now, by definition, if  $A$  occurs then the true bias parameter  $\tau$  is contained in  $\mathcal{T}(\Psi M + \tau, \delta)$  and hence

$$P[\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\} \cap A] \geq 1 - \alpha - \delta.$$

But when  $\tau \in \mathcal{T}(\Psi M + \tau, \delta)$ ,  $a_{min} \leq a(\tau)$  and  $b(\tau) \leq b_{max}$ . It follows that

$$\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\} \cap A \subseteq \{a_{min} \leq \Lambda(\tau) \leq b_{max}\}$$

and therefore

$$1 - \alpha - \delta \leq P[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A] \leq P[\{a_{min} \leq \Lambda(\tau) \leq b_{max}\}]$$

as asserted. □

## References

- Abramson, M., Audet, C., Couture, G., Dennis, Jr., J., Le Digabel, S., Tribes, C., 2013. The NOMAD project. Software available at <http://www.gerad.ca/nomad>.
- Acemoglu, D., Johnson, S., Robinson, J. A., 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91 (5), 1369–1401.
- Andrews, D. W. K., December 1988. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* 4 (3), 458–467.
- Andrews, D. W. K., June 1992. Generic uniform convergence. *Econometric Theory* 8 (2), 241–257.
- Andrews, D. W. K., May 1999. Consistent moment selection procedures for generalized methods of moments estimation. *Econometrica* 67 (3), 543–564.
- Andrews, D. W. K., Lu, B., 2001. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics* 101, 123–164.
- Audet, C., Le Digabel, S., Tribes, C., 2009. NOMAD user guide. Tech. Rep. G-2009-37, Les cahiers du GERAD.  
URL [http://www.gerad.ca/NOMAD/Downloads/user\\_guide.pdf](http://www.gerad.ca/NOMAD/Downloads/user_guide.pdf)
- Berger, R. L., Boos, D. D., September 1994. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 89 (427), 1012–1016.
- Berkowitz, D., Caner, M., Fang, Y., 2012. The validity of instruments revisited. *Journal of Econometrics* 166, 255–266.
- Buckland, S. T., Burnham, K. P., Augustin, N. H., 1997. Model selection: An integral part of inference. *Biometrics* 53 (2), 603–618.
- Carstensen, K., Gundlach, E., 2006. The primacy of institutions reconsidered: Direct income effects of malaria prevalence. *World Bank Economic Review* 20 (3), 309–339.
- Chen, X., Jacho-Chvez, D. T., Linton, O., June 2009. An alternative way of computing efficient instrumental variables estimators, ISE STICERD Research Paper EM/2009/536.  
URL <http://sticerd.lse.ac.uk/dps/em/em536.pdf>
- Cheng, X., Liao, Z., October 2013. Select the valid and relevant moments: An information-based LASSO for GMM with many moments, PIER Working Paper 13-062.  
URL <http://economics.sas.upenn.edu/system/files/13-062.pdf>

- Claeskens, G., Croux, C., Jo, 2006. Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* 62, 972–979.
- Claeskens, G., Hjort, N. L., 2003. The focused information criterion. *Journal of the American Statistical Association* 98 (464), 900–945.
- Claeskens, G., Hjort, N. L., 2008a. Minimizing average risk in regression models. *Econometric Theory* 24, 493–527.
- Claeskens, G., Hjort, N. L., 2008b. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge.
- Conley, T. G., Hansen, C. B., Rossi, P. E., 2012. Plausibly exogenous. *Review of Economics and Statistics* 94 (1), 260–272.
- Demetrescu, M., Hassler, U., Kuzin, V., 2011. Pitfalls of post-model-selection testing: Experimental quantification. *Empirical Economics* 40, 359–372.
- Donald, S. G., Imbens, G. W., Newey, W. K., 2009. Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics* 152, 28–36.
- Donald, S. G., Newey, W. K., September 2001. Choosing the number of instruments. *Econometrica* 69 (5), 1161–1191.
- Easterly, W., Levine, R., 2003. Tropics, germs, and crops: how endowments influence economic development. *Journal of Monetary Economics* 50, 3–39.
- Eddelbuettel, D., 2013. *Seamless R and C++ Integration with Rcpp*. Springer, New York, iISBN 978-1-4614-6867-7.
- Eddelbuettel, D., François, R., 2011. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* 40 (8), 1–18.  
URL <http://www.jstatsoft.org/v40/i08/>
- Eddelbuettel, D., Sanderson, C., March 2014. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis* 71, 1054–1063.  
URL <http://dx.doi.org/10.1016/j.csda.2013.02.005>
- Fox, J., Nie, Z., Byrnes, J., 2014. *sem: Structural Equation Models*. R package version 3.1-4.  
URL <http://CRAN.R-project.org/package=sem>
- Guggenberger, P., 2010. The impact of a hausman pretest on the asymptotic size of a hypothesis test. *Econometric Theory* 26, 369–382.
- Guggenberger, P., 2012. On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econometric Theory* 28, 387–421.

- Guggenberger, P., Kumar, G., 2012. On the size distortion of tests after an overidentifying restrictions pretest. *Journal of Applied Econometrics* 27, 1138–1160.
- Hall, A. R., 2005. *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford.
- Hall, A. R., Peixe, F. P., 2003. A consistent method for the selection of relevant instruments in linear models. *Econometric Reviews* 22, 269–288.
- Hansen, B. E., September 2013. Efficient shrinkage in parametric models, university of Wisconsin.
- Hjort, N. L., Claeskens, G., 2003. Frequentist model average estimators. *Journal of the American Statistical Association* 98 (464), 879–899.
- Hong, H., Preston, B., Shum, M., 2003. Generalized empirical likelihood-based model selection for moment condition models. *Econometric Theory* 19, 923–943.
- Jana, K., 2005. Canonical correlations and instrument selection in econometrics. Ph.D. thesis, North Carolina State University.  
URL <http://www.lib.ncsu.edu/resolver/1840.16/4315>
- Judge, G. G., Mittelhammer, R. C., 2007. Estimation and inference in the case of competing sets of estimating equations. *Journal of Econometrics* 138, 513–531.
- Kabaila, P., 1998. Valid confidence intervals in regressions after variable selection. *Econometric Theory* 14, 463–482.
- Kabaila, P., Leeb, H., 2006. On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* 101 (474), 819–829.
- Kraay, A., 2012. Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach. *Journal of Applied Econometrics* 27 (1), 108–128.
- Kuersteiner, G., Okui, R., March 2010. Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78 (2), 679–718.
- Le Digabel, S., 2011. Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. *ACM Transactions on Mathematical Software* 37 (4), 1–15.
- Leeb, H., Pötscher, B. M., 2005. Model selection and inference: Facts and fiction. *Econometric Theory* 21 (1), 21–59.
- Leeb, H., Pötscher, B. M., 2008. Sparse estimators and the oracle property, or the return of Hodges’ estimator. *Journal of Econometrics* 142, 201–211.



- Leeb, H., Pötscher, B. M., 2009. Model selection. In: Handbook of Financial Time Series. Springer.
- Leeb, H., Pötscher, B. M., September 2014. Testing in the presence of nuisance parameters: Some comments on tests post-model-selection and random critical values, University of Vienna.  
URL <http://arxiv.org/pdf/1209.4543.pdf>
- Liao, Z., November 2013. Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory* 29, 857–904.
- Loh, W.-Y., 1985. A new method for testing separate families of hypotheses. *Journal of the American Statistical Association* 80 (390), 362–368.
- McCloskey, A., October 2012. Bonferroni-based size-correction for nonstandard testing problems, Brown University.  
URL [http://www.econ.brown.edu/fac/adam\\_mccloskey/Research\\_files/McCloskey\\_BBCV.pdf](http://www.econ.brown.edu/fac/adam_mccloskey/Research_files/McCloskey_BBCV.pdf)
- Newey, W. K., 1985. Generalized method of moments specification testing. *Journal of Econometrics* 29, 229–256.
- Newey, W. K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. Vol. IV. Elsevier Science, Ch. 36, pp. 2111–2245.
- Phillips, P. C. B., 1980. The exact distribution of instrumental variables estimators in an equation containing  $n + 1$  endogenous variables. *Econometrica* 48 (4), 861–878.
- Pötscher, B. M., 1991. Effects of model selection on inference. *Econometric Theory* 7, 163–185.
- R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.  
URL <http://www.R-project.org/>
- Racine, J. S., Nie, Z., 2014. crs: Categorical Regression Splines. R package version 0.15-22.  
URL <http://CRAN.R-project.org/package=crs>
- Rodrik, D., Subramanian, A., Trebbi, F., 2004. Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth* 9, 131–165.
- Sachs, J. D., February 2003. Institutions don't rule: Direct effects of geography on per capita income, NBER Working Paper No. 9490.  
URL <http://www.nber.org/papers/w9490>

- Sanderson, C., 2010. Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Tech. rep., NICTA.  
URL [http://arma.sourceforge.net/armadillo\\_nicta\\_2010.pdf](http://arma.sourceforge.net/armadillo_nicta_2010.pdf)
- Schorfheide, F., 2005. VAR forecasting under misspecification. *Journal of Econometrics* 128, 99–136.
- Sharpsteen, C., Bracken, C., 2013. tikzDevice: R Graphics Output in LaTeX Format. R package version 0.7.0.  
URL <http://CRAN.R-project.org/package=tikzDevice>
- Silvapulle, M. J., December 1996. A test in the presence of nuisance parameters. *Journal of the American Statistical Association* 91 (436), 1690–1693.
- Xiao, Z., 2010. The weighted method of moments approach for moment condition models. *Economics Letters* 107, 183–186.
- Yang, Y., 2005. Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika* 92 (4), 937–950.