



# Using invalid instruments on purpose: Focused moment selection and averaging for GMM<sup>☆</sup>



Francis J. DiTraglia

University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, United States

## ARTICLE INFO

### Article history:

Received 9 December 2014

Received in revised form

23 July 2016

Accepted 23 July 2016

Available online 22 August 2016

### JEL classification:

C21

C26

C52

### Keywords:

Moment selection

GMM estimation

Model averaging

Focused Information Criterion

Post-selection estimators

## 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 novel moment selection procedure for GMM – the Focused Moment Selection Criterion (FMSC) – in which moment conditions are chosen not based on their validity but on the MSE of their associated estimator of a user-specified target parameter. The FMSC mimics the situation faced by an applied researcher who begins with a set of relatively mild “baseline” assumptions and must decide whether to impose any of a collection of stronger but more controversial “suspect” assumptions. When the (correctly specified) baseline moment conditions identify the model, the FMSC provides an asymptotically unbiased estimator of asymptotic MSE, allowing us to select over the suspect moment conditions. I go on to 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 simulation-based procedures for inference that can be applied to a variety of formal and informal moment-selection and averaging procedures. Both the FMSC and confidence interval procedures perform well in simulations. I conclude with an empirical example examining the effect of instrument selection on the estimated relationship between malaria and income per capita.

© 2016 Elsevier B.V. All rights reserved.

## 1. Introduction

In finite samples, the addition of a slightly endogenous but highly relevant instrument can reduce 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 selection criterion (FMSC). Rather than selecting only valid moment conditions, the FMSC chooses from a set of potentially mis-specified moment conditions based on the asymptotic mean squared error (AMSE) of their associated GMM estimators of a user-specified scalar target parameter  $\mu$ . 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 suppose that two blocks of moment conditions are available: one that is assumed correctly specified, and another that may not be. This mimics the situation faced by an applied researcher who begins with a “baseline” set of relatively mild maintained assumptions and must decide whether to impose any of a collection of stronger but also more controversial “suspect” assumptions. When the (correctly specified) baseline moment conditions identify the model, the FMSC provides an asymptotically unbiased estimator of AMSE, allowing us select over the suspect moment conditions.<sup>2</sup>

<sup>☆</sup> I thank Aislinn Bohren, Xu Cheng, Gerda Claeskens, Bruce Hansen, Byunghoon Kang, Toru Kitagawa, Hannes Leeb, Adam McCloskey, Serena Ng, Alexei Onatski, Hashem Pesaran, Benedikt Pötscher, Frank Schorfheide, Neil Shephard, Richard J. Smith, Stephen Thiele, Melvyn Weeks, and seminar participants at Brown, Cambridge, Columbia, George Washington, Oxford, Queen Mary, Rutgers, St Andrews, UPenn, Vienna, 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.

E-mail address: [fditra@sas.upenn.edu](mailto:fditra@sas.upenn.edu).

<sup>1</sup> When finite-sample MSE is undefined, AMSE comparisons remain meaningful: see online appendix C.

<sup>2</sup> When this is not the case, it remains possible to use the AMSE framework to carry out a sensitivity analysis: see online appendix B.

The primary goal of the FMSC is to select estimators with low AMSE, but researchers typically wish to report confidence intervals along with parameter estimates. Unfortunately the usual procedures for constructing asymptotic confidence intervals for GMM fail when applied to estimators chosen using a moment selection procedure. A “naïve” 95% confidence interval constructed from the familiar textbook formula will generally under-cover: it will contain the true parameter value far less than 95% of the time because it fails to account for the additional sampling uncertainty that comes from choosing an estimator based on the data. To address the challenging problem of inference post-moment selection, I continue under the local mis-specification framework to derive the limit distribution of “moment average estimators”, data-dependent weighted averages of estimators based on different moment conditions. These estimators are interesting in their own right and include post-moment selection estimators as a special case. I propose two simulation-based procedures for constructing confidence intervals for moment average and post-selection estimators, including the FMSC. First is a “2-Step” confidence interval. I prove that this interval guarantees asymptotically valid inference: the asymptotic coverage of a nominal  $100 \times (1 - \alpha)\%$  interval cannot fall below this level. The price of valid inference, however, is conservatism: the actual coverage of the 2-Step interval typically exceeds its nominal level.<sup>3</sup> As a compromise between the conservatism of the 2-Step interval and the severe under-coverage of the naïve interval I go on to propose a “1-Step” confidence interval. This interval is easier to compute than its 2-Step counterpart and performs well in empirically relevant examples, as I show both theoretically and in simulations below. The 1-Step interval is far shorter than the corresponding 2-Step interval and, while it can under-cover, the magnitude of the size distortion is modest compared to that of the naïve intervals typically reported in applied work.

While my methods apply to general GMM models, 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 transparent form, providing a risk-based justification for the Durbin–Hausman–Wu test, and leading to a novel “minimum-AMSE” averaging estimator that combines OLS and TSLS. The FMSC, averaging estimator, and related confidence interval procedures work well in practice, as I demonstrate in a series of simulation experiments and an empirical example from development economics.

The FMSC and minimum-AMSE averaging estimator considered here are derived for a scalar parameter interest, as this is the most common situation encountered in applied work.<sup>4</sup> As a consequence, Stein-type results do not apply: it is impossible to construct an estimator with uniformly lower risk than the “valid” estimator that uses only the baseline moment conditions. Nevertheless, as my simulation results show, selection and averaging can substantially outperform the valid estimator over large regions of the parameter space, particularly when the “suspect” moment conditions are highly informative and *nearly* correct. This is precisely the situation for which the FMSC is intended.

My approach to moment selection is inspired by the focused information criterion of Claeskens and Hjort (2003), a model selection criterion for maximum likelihood estimation. Like Claeskens

and Hjort (2003), I study AMSE-based selection under mis-specification in a drifting asymptotic framework. In contradistinction, however, I consider moment rather than model selection, and general GMM rather than maximum likelihood estimation. Schorfheide (2005) uses a similar approach to select over forecasts constructed from mis-specified vector autoregression models, developed independently of the FIC. While the use of locally mis-specified moment conditions dates back at least as far as Newey (1985), the idea of using this framework for AMSE-based moment selection, however, is novel.

The existing literature on moment selection primarily aims to consistently select all correctly specified moment conditions while eliminating all invalid ones.<sup>5</sup> This idea begins with Andrews (1999) and is extended by Andrews and Lu (2001) and Hong et al. (2003). More recently, Liao (2013) proposes a shrinkage procedure for consistent GMM moment selection and estimation. In a similar vein, Caner et al. (in press) extend and generalize earlier work by Caner (2009) on LASSO-type model selection for GMM to carry out simultaneous model and moment selection via an adaptive elastic net penalty. Whereas these proposals 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. 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 papers 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 distinguishing feature of the FMSC is focus: rather than a one-size-fits-all criterion, the FMSC is really a method of constructing application-specific moment selection criteria. Consider, for example, a 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.

Like Akaike's Information Criterion (AIC), the FMSC is a *conservative* rather than consistent selection procedure, as it remains random even in the limit. While consistency is a desirable property in many settings, the situation is more complex for model and moment selection: consistent and conservative selection procedures have different strengths, but these strengths cannot be combined (Yang, 2005). The goal of this paper is estimators with low risk. Viewed from this perspective consistent selection criteria suffer from a serious defect: they exhibit unbounded minimax risk (Leeb and Pötscher, 2008). Conservative criteria such as the FMSC do not suffer from this shortcoming. Moreover, as discussed in more detail below, the asymptotics of consistent selection paint a misleading picture of the effects of moment selection on inference. For these reasons, the fact that the FMSC is conservative rather than consistent is an asset in the present context.

Because it studies inference post-moment selection, this paper relates to a vast literature on “pre-test” estimators. For an overview, see Leeb and Pötscher (2005, 2009). 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

<sup>3</sup> This is unavoidable given certain impossibility results concerning post-selection inference. See, e.g. Leeb and Pötscher (2005).

<sup>4</sup> For an extension of the FMSC to vector target parameters, see Online Appendix D.

<sup>5</sup> Under the local mis-specification asymptotics considered below, consistent moment selection criteria simply choose *all* available moment conditions. For details, see Theorem 4.2.

the first paper to treat the problem in general for post-moment selection and moment average estimators in the presence of mis-specification. Some related results appear in Berkowitz et al. (2008, 2012), Guggenberger (2010, 2012); Guggenberger and Kumar (2012), and Caner (2014). While I developed the simulation-based, two-stage confidence interval procedure described below by analogy to a suggestion in Claeskens and Hjort (2008b), Leeb and Pötscher (2014) kindly pointed out that similar constructions have appeared 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 framework within which I study moment averaging is related to 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 computed using valid moment conditions to achieve a minimum variance estimator, I combine estimators computed using potentially invalid conditions with the aim of reducing estimator AMSE. A related idea underlies the combined moments (CM) estimator of Judge and Mittelhammer (2007). For a different approach to combining OLS and TSLS estimators, similar in spirit to the Stein-estimator and developed independently of the work presented here, see Hansen (forthcoming). Cheng et al. (2014) provide related results for Stein-type moment averaging in a GMM context with potentially mis-specified moment conditions.

The results presented here are derived under strong identification and abstract from the many instruments problem. Supplementary simulation results presented in online appendix G.4, however, suggest that the FMSC can nevertheless perform well when the “baseline” assumptions only weakly identify the target parameter. Extending the idea behind the FMSC to allow for weak identification and possibly a large number of moment conditions is a challenging topic that I leave for future research.

The remainder of the paper is organized as follows. Section 2 describes the asymptotic framework and Section 3 derives the FMSC, both in general and for two specific examples: OLS versus TSLS and choosing instrumental variables. Section 4 studies moment average estimators and shows how they can be used to construct valid confidence intervals post-moment selection. Section 5 presents simulation results and Section 6 considers an empirical example from development economics. Proofs appear at the end of the document; computational details and additional material appear in online 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  $g$  is assumed to be correct whereas that associated with  $h$  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 an iid triangular array of random vectors defined on a probability space  $(\mathcal{Y}, \mathcal{F}, \mathbb{P})$  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)  $Z_{ni} \rightarrow_d Z_i$ .

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.  $E[h(Z, \theta_0)] = 0$ . Local mis-specification is *not* intended as a literal description of real-world datasets: it is merely a device that gives an asymptotic bias-variance trade-off that mimics the finite-sample intuition. Moreover, while I work with an iid triangular array for simplicity, the results presented here can be adapted to handle dependent random variables.

### 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  $\hat{\theta}_S$  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>6</sup> Let  $\mathcal{E}_S$  be the  $|S| \times (p+q)$  moment selection matrix corresponding to  $S$ . That is,  $\mathcal{E}_S$  is a matrix of ones and zeros arranged such that  $\mathcal{E}_S f_n(\theta)$  contains only the sample moment conditions used to estimate  $\hat{\theta}_S$ . Thus, the GMM estimator of  $\theta$  based on moment set  $S$  is given by

$$\hat{\theta}_S = \arg \min_{\theta \in \Theta} [\mathcal{E}_S f_n(\theta)]' \tilde{W}_S [\mathcal{E}_S f_n(\theta)]$$

where  $\tilde{W}_S$  is an  $|S| \times |S|$ , positive 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  $g$ . This notation accommodates a wider range of examples, including choosing between OLS and TSLS estimators.

To consider the limit distribution of  $\hat{\theta}_S$ , 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  $\Omega_{hh}$  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  $\mathcal{E}_S$ , define  $F_S = \mathcal{E}_S F$  and  $\Omega_S = \mathcal{E}_S \Omega \mathcal{E}_S'$ . The following high level assumptions are sufficient for the consistency and asymptotic normality of the candidate GMM estimator  $\hat{\theta}_S$ .

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

- (a)  $\theta_0$  lies in the interior of  $\Theta$ , a compact set.
- (b)  $W_S \rightarrow_p W_S$ , a positive definite matrix.

<sup>6</sup> Identifying  $\tau$  requires further assumptions, as discussed in Section 2.3.

- (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 [Assumption 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 online appendix E. 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\(a\)–\(e\)](#),  $\hat{\theta}_S \rightarrow_p \theta_0$ .

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

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

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  $\hat{\theta}_S$  is consistent for  $\theta_0$  under local mis-specification. Unless  $S$  excludes all of the moment conditions contained in  $h$ , however,  $\hat{\theta}_S$  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  $\hat{\theta}_S$  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>7</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  $\theta' \neq \theta_0$ ,  $|S_{max}| > |S'|$ .

[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 very simple settings. 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 \(2013\)](#).<sup>8</sup>

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

$$\hat{\theta}_v = \arg \min_{\theta \in \Theta} g_n(\theta)' \tilde{W}_v g_n(\theta).$$

We call this the “valid estimator” and assume that it satisfies all the conditions of [Assumption 2.2](#). Note that this implies  $p \geq r$ .

[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. In the context of the present paper, [Assumption 2.4](#) is meant to represent a situation in which an applied researcher chooses between two groups of assumptions. The  $g$ -block contains the “baseline” assumptions while the  $h$ -block contains a set of stronger, more controversial “suspect” 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.

## 3. The focused moment selection criterion

### 3.1. The general case

The FMSC chooses among the potentially invalid moment conditions contained in  $h$  based on the estimator AMSE of a user-specified scalar target parameter.<sup>9</sup> 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  $\theta_S$  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  $\hat{\mu}_S$ .

<sup>8</sup> For a discussion of why [Assumption 2.4](#) is necessary and how to proceed when it fails, see online appendix B.

<sup>9</sup> Although I focus on the case of a scalar target parameter in the body of the paper, the same idea can be applied to a vector of target parameters. For details see online appendix D.

<sup>7</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  $\tau \rightarrow \infty$  sufficiently fast.



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

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

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

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

For the valid estimator  $\hat{\theta}_v$  we have  $K_v = [G'W_vG]^{-1}G'W_v$  and  $\mathcal{E}_v = [\mathbf{I}_p \quad \mathbf{0}_{p \times q}]$ . 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 extent and direction of which depend both on  $K_S$  and  $\nabla_{\theta}\mu(\theta_0)$ . The setting considered here, however, is one in which using moment conditions from  $h$  in estimation will reduce the asymptotic variance. In the nested case, where moment conditions from  $h$  are added to those of  $g$ , this follows automatically. 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 misspecification, because all moment conditions are correctly specified in the limit. In non-nested examples, for example when  $h$  contains OLS moment conditions and  $g$  contains IV moment conditions, however, this result does not apply because one would use  $h$  instead of  $g$ . In such examples, one must establish an analogous ordering of asymptotic variances by direct calculation, as I do below for the OLS versus IV example.

Using this framework for moment selection requires estimators of the unknown quantities:  $\theta_0$ ,  $K_S$ ,  $\Omega$ , and  $\tau$ . Under local misspecification, the estimator of  $\theta$  under any moment set is consistent. A natural estimator is  $\hat{\theta}_v$ , although there are other possibilities. Recall that  $K_S = [F_S'W_SF_S]^{-1}F_S'W_SF_S$ . Because it is simply the selection matrix defining moment set  $S$ ,  $\mathcal{E}_S$  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 misspecification, although the precise form depends on the situation.<sup>10</sup> The only remaining unknown is  $\tau$ . Local misspecification 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 misspecified 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

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

where  $K_v$  is defined in Corollary 2.1 and  $\mathbf{I}_q$  denotes the  $(q \times q)$  identity matrix. 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

$$\text{FMSC}_n(S) = \nabla_{\theta}\mu(\hat{\theta})'K_S\mathcal{E}_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \hat{\tau}\hat{\tau}' - \hat{\Psi}\hat{\Omega}\hat{\Psi}' \end{bmatrix} + \hat{\Omega} \right\} \mathcal{E}_S'K_S'\nabla_{\theta}\mu(\hat{\theta}) \quad (1)$$

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 Eq. (1), that is  $S_{\text{FMSC}}^* = \arg \min_{S \in \mathcal{S}} \text{FMSC}_n(S)$ .

In summary, the FMSC aims to choose the moment conditions that provide the lowest risk estimator of a target parameter  $\mu$  where risk is defined as MSE.<sup>11</sup> Because finite-sample MSE is unavailable, AMSE in a local-to-zero asymptotic framework serves in its stead. Since no consistent estimator of AMSE exists in this setting, FMSC uses an asymptotically unbiased estimator. This is the same idea 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).

### 3.2. OLS versus TSLS example

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  $y$ . The intuition is straightforward: because TSLS is a high-variance estimator, OLS will have a lower mean-squared error provided that  $x$  is not too endogenous.<sup>12</sup> To keep the presentation transparent, I work within an iid, homoskedastic setting for this example and assume, without loss of generality, that there are no exogenous regressors.<sup>13</sup> Equivalently we may suppose that any exogenous regressors, including a constant, have been “projected out”. Low-level sufficient conditions for all of the results in this section appear in Assumption E.1 of online appendix E. The data generating process is

$$y_{ni} = \beta x_{ni} + \epsilon_{ni} \quad (2)$$

$$x_{ni} = \mathbf{z}_{ni}'\pi + v_{ni} \quad (3)$$

where  $\beta$  and  $\pi$  are unknown constants,  $\mathbf{z}_{ni}$  is a vector of exogenous and relevant instruments,  $x_{ni}$  is the endogenous regressor,  $y_{ni}$  is the outcome of interest, and  $\epsilon_{ni}$ ,  $v_{ni}$  are unobservable error terms. All random variables in this system are mean zero, or equivalently all constant terms have been projected out. Stacking observations

<sup>11</sup> One could choose a different risk function and proceed similarly, although I do not consider this idea further below. See, e.g., Claeskens and Croux (2006) and Claeskens and Hjort (2008a).

<sup>12</sup> Because the moments of the TSLS estimator only exist up to the order of overidentification (Phillips, 1980; Kinal, 1980) mean-squared error should be understood to refer to “trimmed” mean-squared error when the number of instruments is two or fewer. For details, see online appendix C.

<sup>13</sup> The homoskedasticity assumption concerns the limit random variables: under local misspecification there will be heteroskedasticity for fixed  $n$ . See Assumption E.1 in online appendix E for details.

<sup>10</sup> See Sections 3.2 and 3.3 for discussion of this point for the two running examples.

in the usual way, the estimators under consideration are  $\hat{\beta}_{OLS} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$  and  $\tilde{\beta}_{TSLs} = (\mathbf{x}'P_Z\mathbf{x})^{-1}\mathbf{x}'P_Z\mathbf{y}$  where we define  $P_Z = Z(Z'Z)^{-1}Z'$ .

**Theorem 3.2** (OLS and TSLs Limit Distributions). *Let  $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni})$  be a triangular array of random variables such that  $E[\mathbf{z}_{ni}\epsilon_{ni}] = \mathbf{0}$ ,  $E[\mathbf{z}_{ni}v_{ni}] = \mathbf{0}$ , and  $E[\epsilon_{ni}v_{ni}] = \tau/\sqrt{n}$  for all  $n$ . Then, under standard regularity conditions, e.g. Assumption E.1 in online appendix E,*

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

where  $\sigma_x^2 = \gamma^2 + \sigma_v^2$ ,  $\gamma^2 = \pi'Q\pi$ ,  $E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \rightarrow Q$ ,  $E[v_{ni}^2] \rightarrow \sigma_v^2$ , and  $E[\epsilon_{ni}^2] \rightarrow \sigma_\epsilon^2$  as  $n \rightarrow \infty$ .

We see immediately that, as expected, 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  $\tau$ . Thus,

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

and rearranging, we see that the AMSE of the OLS estimator is strictly less than that of the TSLs estimator whenever  $\tau^2 < \sigma_x^2\sigma_\epsilon^2/\gamma^2$ . To estimate the unknowns required to turn this inequality into a moment selection procedure, I set

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

and define

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

Under local mis-specification each of these estimators is consistent for its population counterpart.<sup>14</sup> 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 3.3.** *Let  $\hat{\tau} = n^{-1/2}\mathbf{x}'(\mathbf{y} - \mathbf{x}\tilde{\beta}_{TSLs})$ . Then, under the conditions of Theorem 3.2,*

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

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  $\hat{V} = \hat{\sigma}_v^2\hat{\sigma}_\epsilon^2\hat{\sigma}_x^2/\hat{\gamma}^2$ . The quantity  $\hat{T}_{FMSC}$  looks very much like a test statistic and indeed it can be viewed as such. By Theorem 3.3 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, the FMSC is numerically equivalent to using OLS unless the DHW test rejects at the 16% level.

**Theorem 3.4.** *Under the conditions of Theorem 3.2, FMSC selection between the OLS and TSLs estimators is equivalent to a Durbin–Hausman–Wu pre-test with a critical value of 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 idea behind the familiar DHW test to more general GMM environments.<sup>15</sup>

### 3.3. Choosing instrumental variables example

The OLS versus TSLs example is really a special case of instrument selection: if  $x$  is exogenous, it is clearly “its own best instrument”. Viewed from this perspective, the FMSC amounts to trading off endogeneity against instrument strength. I now consider instrument selection in general for linear GMM estimators in an iid setting. Consider the model:

$$y_{ni} = \mathbf{x}_{ni}'\beta + \epsilon_{ni} \quad (4)$$

$$\mathbf{x}_{ni} = \Pi_1'\mathbf{z}_{ni}^{(1)} + \Pi_2'\mathbf{z}_{ni}^{(2)} + \mathbf{v}_{ni} \quad (5)$$

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  $\beta$ ,  $p \times r$  matrix  $\Pi_1$ , and  $q \times r$  matrix  $\Pi_2$  contain unknown constants. Stacking observations in the usual way, we can write the system in matrix form as  $\mathbf{y} = X\beta + \epsilon$  and  $X = Z\Pi + V$ , where  $Z = (Z_1, Z_2)$  and  $\Pi = (\Pi_1', \Pi_2')'$ .

In this example, the idea 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 they 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. Accordingly, the FMSC attempts to trade off a small increase in bias from using a slightly endogenous instrument against a larger decrease in variance from increased instrument strength. To this end, consider a general linear GMM estimator of the form

$$\hat{\beta}_S = (X'S\tilde{W}_S Z_S'X)^{-1}X'S\tilde{W}_S Z_S'\mathbf{y}$$

where  $S$  indexes the instruments used in estimation,  $Z_S' = \mathcal{E}_S Z'$  is the matrix containing only those instruments included in  $S$ ,  $|S|$  is the number of instruments used in estimation and  $\tilde{W}_S$  is an  $|S| \times |S|$  positive definite weighting matrix.

**Theorem 3.5** (Choosing IVs Limit Distribution). *Let  $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni})$  be a triangular array of random variables such that  $E[\mathbf{z}_{ni}\epsilon_{ni}] = \mathbf{0}$ ,  $E[\mathbf{z}_{ni}v_{ni}] = \mathbf{0}$ , and  $E[\epsilon_{ni}v_{ni}] = \tau/\sqrt{n}$  for all  $n$ . Suppose further that  $\tilde{W}_S \rightarrow_p W_S > 0$ . Then, under standard regularity conditions, e.g. Assumption E.2 in online appendix E,*

$$\sqrt{n}(\hat{\beta}_S - \beta) \xrightarrow{d} -K_S \mathcal{E}_S \begin{bmatrix} \mathbf{0} \\ \tau \end{bmatrix} + M$$

<sup>14</sup> Note that the FMSC in this example, characterized in Theorem 3.4, chooses between OLS and IV to minimize estimator AMSE. If one wishes to carry out inference post-selection one must contend with the size distortions of the familiar “textbook” confidence interval procedure, as pointed out by Cuggenberger (2010). I discuss this point extensively below in Section 4 and propose possible remedies.

<sup>15</sup> While using the OLS residuals to estimate  $\sigma_\epsilon^2$  also provides a consistent estimate under local mis-specification, the estimator based on the TSLs residuals should be more robust.

where

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

$M \sim N(\mathbf{0}, \Omega)$ ,  $Q_S = Q \mathcal{E}_S'$ ,  $E[\mathbf{z}_{ni} \mathbf{z}_{ni}'] \rightarrow Q$  and  $E[\epsilon_{ni}^2 \mathbf{z}_{ni} \mathbf{z}_{ni}'] \rightarrow \Omega$  as  $n \rightarrow \infty$ .

To implement the FMSC for this example, we simply need to specialize Eq. (1). To simplify the notation, let

$$\mathcal{E}_1 = [I_p \quad 0_{p \times q}], \quad \mathcal{E}_2 = [0_{q \times p} \quad I_q] \quad (6)$$

where  $0_{p \times q}$  denotes a  $p \times q$  matrix of zeros and  $I_q$  denotes the  $q \times q$  identity matrix. Using this convention,  $Z_1 = Z \mathcal{E}_1'$  and  $Z_2 = Z \mathcal{E}_2'$ . In this example the valid estimator, defined in Assumption 2.4, is given by

$$\hat{\beta}_v = (X' Z_1 \tilde{W}_v Z_1' X)^{-1} X' Z_1 \tilde{W}_v Z_1' y \quad (7)$$

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

$$-\hat{K}_S = n (X' Z \mathcal{E}_S' \tilde{W}_S \mathcal{E}_S' X)^{-1} X' Z \mathcal{E}_S' \tilde{W}_S$$

is the natural consistent estimator of  $-K_S$  in this setting.<sup>16</sup> Since  $\mathcal{E}_S$  is known, the only remaining quantities from Eq. (1) are  $\hat{\tau}$ ,  $\hat{\Psi}$  and  $\hat{\Omega}$ . The following result specializes Theorem 3.1 to the present example.

**Theorem 3.6.** Let  $\hat{\tau} = n^{-1/2} Z_2' (y - X \hat{\beta}_v)$  where  $\hat{\beta}_v$  is as defined in Eq. (7). Under the conditions of Theorem 3.5 we have  $\hat{\tau} \rightarrow_d \tau + \Psi M$  where  $M$  is defined in Theorem 3.5,

$$\Psi = [-\mathcal{E}_2 Q \Pi K_v \quad I_q]$$

$$-K_v = (\Pi' Q \mathcal{E}_1' W_v \mathcal{E}_1 Q' \Pi)^{-1} \Pi' Q \mathcal{E}_1' W_v$$

$W_v$  is the probability limit of the weighting matrix from Eq. (7),  $I_q$  is the  $q \times q$  identity matrix,  $\mathcal{E}_1$  is defined in Eq. (6), and  $E[\mathbf{z}_{ni} \mathbf{z}_{ni}'] \rightarrow Q$ .

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

$$\hat{\Psi} = [-n^{-1} Z_2' X (-\hat{K}_v) \quad I_q], \quad -\hat{K}_v = n (X' Z_1 \tilde{W}_v Z_1' X)^{-1} X' Z_1 \tilde{W}_v.$$

All that remains before substituting values into Eq. (1) is to estimate  $\Omega$ . In the simulation and empirical examples discussed below I examine the TSLS estimator, that is  $W_S = (\mathcal{E}_S' Z' Z \mathcal{E}_S)^{-1}$ , and estimate  $\Omega$  as follows. For all specifications except the valid estimator  $\hat{\beta}_v$ , I employ the centered, heteroskedasticity-consistent estimator

$$\hat{\Omega}_S = \frac{1}{n} \sum_{i=1}^n u_i(\hat{\beta}_S)^2 \mathbf{z}_{iS} \mathbf{z}_{iS}' - \left( \frac{1}{n} \sum_{i=1}^n u_i(\hat{\beta}_S) \mathbf{z}_{iS} \right) \left( \frac{1}{n} \sum_{i=1}^n u_i(\hat{\beta}_S) \mathbf{z}_{iS}' \right) \quad (8)$$

where  $u_i(\beta) = y_i - \mathbf{x}_i' \beta$ ,  $\hat{\beta}_S = (X' Z_S (Z_S' Z_S)^{-1} Z_S' X)^{-1} X' Z_S (Z_S' Z_S)^{-1} Z_S' y$ ,  $\mathbf{z}_{iS} = \mathcal{E}_S \mathbf{z}_i$  and  $Z_S' = \mathcal{E}_S' Z'$ . 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  $\hat{\beta}_v$  has no asymptotic bias, the AMSE of any target parameter based on this estimator equals its asymptotic variance. Accordingly, I use

$$\tilde{\Omega}_{11} = n^{-1} \sum_{i=1}^n u_i(\hat{\beta}_v)^2 \mathbf{z}_{1i} \mathbf{z}_{1i}' \quad (9)$$

rather than the  $(p \times p)$  upper left sub-matrix of  $\hat{\Omega}$  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.

#### 4. Moment averaging and post-selection estimators

Because it is constructed from  $\hat{\tau}$ , the FMSC is a random variable, even in the limit. Combining Corollary 3.2 with Eq. (1) gives the following.

**Corollary 4.1** (Limit Distribution of FMSC). Under Assumptions 2.1, 2.2 and 2.4, we have  $FMSC_n(S) \rightarrow_d FMSC_S(\tau, M)$ , where  $B(\tau, M) = (\Psi M + \tau)(\Psi M + \tau)' - \Psi \Omega \Psi'$  and

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

This corollary implies that the FMSC is a “conservative” rather than “consistent” selection procedure. This lack of consistency is 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. 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.<sup>17</sup> 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.

There are two main problems with applying “textbook” confidence intervals post-moment selection. First is model selection uncertainty: 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 model must be too short. Second, textbook confidence intervals ignore the fact that selection is carried out over potentially invalid moment conditions. Even if our goal were to consistently eliminate such 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. Because of this, our intervals will be incorrectly centered. Accounting for these two effects requires a limit theory that accommodates mixture distributions: post-selection estimators are randomly-weighted averages of the individual candidate estimators. Because they choose a single candidate with probability approaching one in the limit, consistent selection procedures make it impossible to represent this phenomenon. In contrast, conservative selection procedures remain random even as the sample size goes to infinity, allowing us to derive a mixture-of-normals limit distribution and, ultimately, to carry out valid inference post-moment selection. In the remainder of this section, I derive the asymptotic distribution of generic “moment average” estimators and use them to propose simulation-based procedures for post-moment selection inference. For certain examples it is possible to analytically characterize the limit distribution of a post-FMSC estimator without resorting to simulation-based methods. I explore this possibility in detail for my two running examples: OLS versus TSLS and choosing instrumental variables. I also briefly consider a minimum-AMSE averaging estimator that combines OLS and TSLS.

<sup>16</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.

<sup>17</sup> This fact is readily apparent from the results of the simulation study from Section 5.2: the consistent criteria, GMM-BIC and HQ, have the highest worst-case RMSE, while the conservative criteria, FMSC and GMM-AIC, have the lowest.

#### 4.1. Moment average estimators

A generic moment average estimator takes the form

$$\hat{\mu} = \sum_{S \in \mathcal{S}} \hat{\omega}_S \hat{\mu}_S \quad (10)$$

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  $\hat{\omega}_S$  is an indicator, taking on the value one at the moment set that minimizes some moment selection criterion,  $\hat{\mu}$  is a post-moment selection estimator. To characterize the limit distribution of  $\hat{\mu}$ , I impose the following mild conditions on  $\hat{\omega}_S$ , requiring that they sum to one and are “well-behaved” in the limit so that I may apply the continuous mapping theorem.

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

- (a)  $\sum_{S \in \mathcal{S}} \hat{\omega}_S = 1$ , almost surely.
- (b) For each  $S \in \mathcal{S}$ ,  $\hat{\omega}_S \rightarrow_d \varphi_S(\tau, M)$ , a function of  $\tau$ ,  $M$  and consistently estimable constants with at most countably many discontinuities.

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

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

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  $\hat{\mu}$  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 weights based on 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  $J_n(S) = n[\Xi_S f_n(\hat{\theta}_S)]' \hat{\Omega}_S^{-1} [\Xi_S f_n(\hat{\theta}_S)]$  where  $\hat{\Omega}_S^{-1}$  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  $\Omega_S^{-1/2} F_S$ .

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

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

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  $S \neq S'$ , we have

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

satisfying Assumption 4.1(b). Thus, post-selection estimators based on the FMSC, a downward  $J$ -test procedure, or the GMM moment selection criteria of Andrews (1999) all fall within the ambit of 4.2. The consistent criteria of Andrews (1999), however, are not particularly interesting under local mis-specification.<sup>18</sup> Intuitively, because they aim to select all valid moment conditions w.p.a.1, we would expect that under Assumption 2.1 they choose the full moment set in the limit. The following result shows that this intuition is correct.<sup>19</sup>

**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.

#### 4.2. Digression: Minimum-AMSE averaging for OLS and TSLS

When competing moment sets have similar criterion values in the population, sampling variation can be magnified in the selected estimator. This motivates the idea of averaging estimators based on different moment conditions rather than selecting them. To illustrate this idea, I now briefly revisit the OLS versus TSLS example from Section 3.2 and derive an AMSE-optimal weighted average of the two estimators. Let  $\tilde{\beta}(\omega)$  be a convex combination of the OLS and TSLS estimators, namely

$$\tilde{\beta}(\omega) = \omega \hat{\beta}_{OLS} + (1 - \omega) \tilde{\beta}_{TSLS} \quad (11)$$

where  $\omega \in [0, 1]$  is the weight given to the OLS estimator.

**Theorem 4.3.** Under the conditions of Theorem 3.2, the AMSE of the weighted-average estimator  $\sqrt{n}[\tilde{\beta}(\omega) - \beta]$  is minimized over  $\omega \in [0, 1]$  by taking  $\omega = \omega^*$  where

$$\omega^* = \left[ 1 + \frac{\tau^2 / \sigma_x^4}{\sigma_\epsilon^2 (1/\gamma^2 - 1/\sigma_x^2)} \right]^{-1} = \left[ 1 + \frac{ABIAS(OLS)^2}{AVAR(TSLS) - AVAR(OLS)} \right]^{-1}.$$

The preceding result 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  $z$  as measured by  $\gamma$ . All else equal, the stronger the instruments, the less weight we should give to OLS. To operationalize the AMSE-optimal averaging estimator suggested from Theorem 4.3, I propose the plug-in estimator

$$\hat{\beta}_{AVG}^* = \hat{\omega}^* \hat{\beta}_{OLS} + (1 - \hat{\omega}^*) \tilde{\beta}_{TSLS} \quad (12)$$

<sup>18</sup> For more discussion of these criteria, see Section 5.2.

<sup>19</sup> This result is a special case of a more general phenomenon: consistent selection procedures cannot detect model violations of order  $O(n^{-1/2})$ .



where

$$\hat{\omega}^* = \left[ 1 + \frac{\max \{0, (\hat{\tau}^2 - \hat{\sigma}_\epsilon^2 \hat{\sigma}_x^2 (\hat{\sigma}_x^2 / \hat{\gamma}^2 - 1)) / \hat{\sigma}_x^4\}}{\hat{\sigma}_\epsilon^2 (1 / \hat{\gamma}^2 - 1 / \hat{\sigma}_x^2)} \right]^{-1}. \quad (13)$$

This expression employs the same consistent estimators of  $\sigma_x^2$ ,  $\gamma$  and  $\sigma_\epsilon$  as the FMSC expressions from Section 3.2. To ensure that  $\hat{\omega}^*$  lies in the interval  $[0, 1]$ , however, I use a *positive part* estimator for  $\tau^2$ , namely  $\max\{0, \hat{\tau}^2 - \hat{V}\}$  rather than  $\hat{\tau}^2 - \hat{V}$ .<sup>20</sup> In the following section I show how one can construct confidence intervals for  $\beta^*$  and related estimators.

#### 4.3. Inference for moment-average estimators

Suppose that  $K_S$ ,  $\varphi_S$ ,  $\theta_0$ ,  $\Omega$  and  $\tau$  were all 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. This is the basic intuition that I use to devise inference procedures for moment-average and post-selection estimators.

To operationalize this idea, first consider how we would proceed if we knew *only* the value of  $\tau$ . While  $K_S$ ,  $\theta_0$ , and  $\Omega$  are unknown this presents only a minor difficulty: in their place we can simply substitute the consistent estimators that appeared in the expression for the FMSC above. To estimate  $\varphi_S$ , we first need to derive the limit distribution of  $\hat{\omega}_S$ , the data-based weights specified by the user. As an example, consider the case of moment selection based on the FMSC. Here  $\hat{\omega}_S$  is simply the indicator function

$$\hat{\omega}_S = \mathbf{1} \left\{ \text{FMSC}_n(S) = \min_{S' \in \mathcal{S}} \text{FMSC}_n(S') \right\}. \quad (14)$$

Substituting estimators of  $\Omega$ ,  $K_S$  and  $\theta_0$  into  $\text{FMSC}_S(\tau, M)$ , defined in Corollary 4.1, gives

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

where  $\hat{B}(\tau, M) = (\hat{\Psi}M + \tau)(\hat{\Psi}M + \tau)' - \hat{\Psi}\hat{\Omega}\hat{\Psi}$ . Combining this with Eq. (14),

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

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. Continuing to assume for the moment that  $\tau$  is known, consider the following algorithm:

#### Algorithm 4.1 (Simulation-Based CI for $\hat{\mu}$ Given $\tau$ ).

1. Generate  $J$  independent draws  $M_j \sim N_{p+q}(0, \hat{\Omega})$ .
2. Set  $\Lambda_j(\tau) = -\nabla_\theta \mu(\hat{\theta})' [\sum_{S \in \mathcal{S}} \hat{\varphi}_S(\tau, M_j) \hat{K}_S \hat{\mathcal{E}}_S] (M_j + \tau)$ .
3. Using  $\{\Lambda_j(\tau)\}_{j=1}^J$ , calculate  $\hat{a}(\tau)$ ,  $\hat{b}(\tau)$  such that  $P\{\hat{a}(\tau) \leq \Lambda(\tau) \leq \hat{b}(\tau)\} = 1 - \alpha$ .
4. Define the interval  $\text{CI}_{\text{sim}} = [\hat{\mu} - \hat{b}(\tau)/\sqrt{n}, \hat{\mu} - \hat{a}(\tau)/\sqrt{n}]$ .

Given knowledge of  $\tau$ , Algorithm 4.1 yields valid inference for  $\mu$ . The problem, of course, is that  $\tau$  is unknown and cannot even be consistently estimated. One idea would be to substitute the asymptotically unbiased estimator  $\hat{\tau}$  from 3.1 in place of the unknown  $\tau$ . This gives rise to a procedure that I call the “1-Step” confidence interval:

**Algorithm 4.2 (1-Step CI).** Carry out of Algorithm 4.1 with  $\tau$  set equal to the estimator  $\hat{\tau}$  from Theorem 3.1, yielding  $\hat{\text{CI}}_1 = [\hat{\mu} - \hat{b}(\hat{\tau})/\sqrt{n}, \hat{\mu} - \hat{a}(\hat{\tau})/\sqrt{n}]$ .

The 1-Step interval defined in Algorithm 4.2 is conceptually simple, easy to compute, and can perform well in practice, as I explore below. But as it fails to account for sampling uncertainty in  $\hat{\tau}$ , it does *not* necessarily yield asymptotically valid inference for  $\mu$ . Fully valid inference requires the addition of a second step to the algorithm and comes at a cost: conservative rather than exact inference. In particular, the following procedure is guaranteed to yield an interval with asymptotic coverage probability of *at least*  $(1 - \alpha - \delta) \times 100\%$ .

#### Algorithm 4.3 (2-Step CI).

1. Construct a  $(1 - \delta) \times 100\%$  confidence region  $\mathcal{T}$  for  $\tau$  using Theorem 3.1.
2. For each  $\tau^* \in \mathcal{T}$  carry out Algorithm 4.1, yielding a  $(1 - \alpha) \times 100\%$  confidence interval  $[\hat{a}(\tau^*), \hat{b}(\tau^*)]$  for  $\Lambda(\tau^*)$ .
3. Set  $\hat{a}_{\min} = \min_{\tau^* \in \mathcal{T}} \hat{a}(\tau^*)$  and  $\hat{b}_{\max} = \max_{\tau^* \in \mathcal{T}} \hat{b}(\tau^*)$ .
4. Construct the interval  $\hat{\text{CI}}_2 = [\hat{\mu} - \hat{b}_{\max}/\sqrt{n}, \hat{\mu} - \hat{a}_{\min}/\sqrt{n}]$ .

**Theorem 4.4 (2-Step CI).** 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  $\Delta_n(\hat{\tau}, \tau^*) = (\hat{\tau} - \tau^*)' (\hat{\Psi} \hat{\Omega} \hat{\Psi}')^{-1} (\hat{\tau} - \tau^*)$  and  $\mathcal{T}(\hat{\tau}, \delta) = \{\tau^*: \Delta_n(\hat{\tau}, \tau^*) \leq \chi_q^2(\delta)\}$  where  $\chi_q^2(\delta)$  denotes the  $1 - \delta$  quantile of a  $\chi^2$  distribution with  $q$  degrees of freedom. Then, the interval  $\text{CI}_2$  defined in Algorithm 4.3 has asymptotic coverage probability no less than  $1 - (\alpha + \delta)$  as  $J, n \rightarrow \infty$ .

#### 4.4. A special case of post-FMSC inference

The preceding section presented two confidence intervals that account for the effects of moment selection on subsequent inference. The 1-Step interval is intuitive and computationally straightforward but lacks theoretical guarantees, while the 2-Step interval guarantees asymptotically valid inference at the cost of greater computational complexity and conservatism. To better understand these methods and the trade-offs involved in deciding between them, I now specialize them to the two examples of FMSC selection that appear in the simulation studies described below. The structure of these examples allows us to bypass Algorithm 4.3 and characterize the asymptotic properties of various proposals for post-FMSC without resorting to Monte Carlo simulations. Because this section presents asymptotic results, I treat any consistently estimable quantity that appears in a limit distribution as known.

In both the OLS versus IV example from Section 3.2 and the slightly simplified version of the choosing instrumental variables example implemented in Section 5.2, the post-FMSC estimator  $\hat{\beta}_{\text{FMSC}}$  converges to a very convenient limit experiment.<sup>21</sup> In particular,

$$\sqrt{n}(\hat{\beta}_{\text{FMSC}} - \beta) \xrightarrow{d} \mathbf{1}\{|T| < \sigma\sqrt{2}\} U + \mathbf{1}\{|T| \geq \sigma\sqrt{2}\} V \quad (15)$$

with

$$T = \sigma Z_1 + \tau, \quad U = \eta Z_2 + c\tau, \quad V = \eta Z_2 - c\sigma Z_1 \quad (16)$$

where  $Z_1, Z_2$  are independent standard normal random variables,  $\eta, \sigma$  and  $c$  are consistently estimable constants, and  $\tau$  is the

<sup>20</sup> While  $\hat{\tau}^2 - \hat{V}$  is an asymptotically unbiased estimator of  $\tau^2$  it can be negative.

<sup>21</sup> The simplified version of the choosing instrumental variables example considers a single potentially endogenous instrument and imposes homoskedasticity. For more details see Section 5.2 and online appendix F.1.

local mis-specification parameter. This representation allows us to tabulate the asymptotic distribution,  $F_{FMSC}$  as follows:

$$F_{FMSC}(x) = G(x) + H_1(x) + H_2(x) \quad (17)$$

$$G(x) = \Phi\left(\frac{x - c\tau}{\eta}\right) \left[ \Phi(\sqrt{2} - \tau/\sigma) - \Phi(-\sqrt{2} - \tau/\sigma) \right] \quad (18)$$

$$H_1(x) = \frac{1}{\sigma} \int_{-\infty}^{-\sigma\sqrt{2}-\tau} \Phi\left(\frac{x + ct}{\eta}\right) \varphi(t/\sigma) dt \quad (19)$$

$$H_2(x) = \frac{1}{\sigma} \int_{\sigma\sqrt{2}-\tau}^{+\infty} \Phi\left(\frac{x + ct}{\eta}\right) \varphi(t/\sigma) dt \quad (20)$$

where  $\Phi$  is the CDF and  $\varphi$  the pdf of a standard normal random variable. Note that the limit distribution of the post-FMSC distribution depends on  $\tau$  in addition to the consistently estimable quantities  $\sigma$ ,  $\eta$ ,  $c$  although I suppress this dependence to simplify the notation. While these expressions lack a closed form  $G$ ,  $H_1$  and  $H_2$  are easy to compute, allowing us to calculate both  $F_{FMSC}$  and the corresponding quantile function  $Q_{FMSC}$ .<sup>22</sup>

The ability to compute  $F_{FMSC}$  and  $Q_{FMSC}$  allows us to answer a number of important questions about post-FMSC inference. First, suppose that we were to carry out FMSC selection and then construct a  $(1 - \alpha) \times 100\%$  confidence interval *conditional* in the selected estimator, completely ignoring the effects of the moment selection step. What would be the resulting asymptotic coverage probability and width of such a “naïve” confidence interval procedure? Using calculations similar to those used above in the expression for  $F_{FMSC}$ , we find that the coverage probability of this naïve interval is given by

$$\begin{aligned} CP_{Naive}(\alpha) &= G(u_\alpha) - G(-u_\alpha) + H_1(\ell_\alpha) - H_2(-\ell_\alpha) \\ &\quad + H_2(\ell_\alpha) - H_2(-\ell_\alpha) \\ u_\alpha &= \Phi(1 - \alpha/2) \eta \\ \ell_\alpha &= \Phi(1 - \alpha/2) \sqrt{\eta^2 + c^2\sigma^2} \end{aligned}$$

where  $G$ ,  $H_1$ ,  $H_2$  are as defined in Eqs. (19)–(20). And since the width of this naïve CI equals that of the textbook interval for  $\hat{\beta}$  when  $|\hat{\tau}| < \sigma\sqrt{2}$  and that of the textbook interval for  $\tilde{\beta}$  otherwise, we have

$$\frac{E[\text{Width}_{Naive}(\alpha)]}{\text{Width}_{Valid}(\alpha)} = 1 + \left[ \Phi(\sqrt{2} - \tau/\sigma) - \Phi(-\sqrt{2} - \tau/\sigma) \right] \times \left( \sqrt{\frac{\eta^2}{\eta^2 + c^2\sigma^2}} - 1 \right)$$

where  $\text{Width}_{Valid}(\alpha)$  is the width of a standard, textbook confidence interval for  $\beta$ .

To evaluate these expressions we need values for  $c$ ,  $\eta^2$ ,  $\sigma^2$  and  $\tau$ . For the remainder of this section I will consider the parameter values that correspond to the simulation experiments presented below in Section 5. For the OLS versus TSLS example we have  $c = 1$ ,  $\eta^2 = 1$  and  $\sigma^2 = (1 - \pi^2)/\pi^2$  where  $\pi^2$  denotes the population first-stage  $R$ -squared for the TSLS estimator. For the choosing IVs example we have  $c = \gamma/(\gamma^2 + 1/9)$ ,  $\eta^2 = 1/(\gamma^2 + 1/9)$  and  $\sigma^2 = 1 + 9\gamma^2$  where  $\gamma^2$  is the increase in the population first-stage  $R$ -squared of the TSLS estimator from adding  $w$  to the instrument set.<sup>23</sup>

Table 1 presents the asymptotic coverage probability and Table 2 the expected relative width of the naïve confidence interval

procedure for a variety of values of  $\tau$  and  $\alpha$  for each of the two examples. For the OLS versus TSLS example, I allow  $\pi^2$  to vary while for the choosing IVs example I allow  $\gamma^2$  to vary. Note that the relative expected width does not depend on  $\alpha$ . In terms of coverage probability, the naïve interval performs very poorly: in some regions of the parameter space the actual coverage is very close to the nominal level, while in others it is far lower. These striking size distortions, which echo the findings of Guggenberger (2010, 2012), provide a strong argument against the use of the naïve interval. Its attraction, of course, is width: the naïve interval can be dramatically shorter than the corresponding “textbook” confidence interval for the valid estimator.

Is there any way to construct a post-FMSC confidence interval that does not suffer from the egregious size distortions of the naïve interval but is still shorter than the textbook interval for the valid estimator? As a first step towards answering this question, Table 3 presents the relative width of the shortest possible *infeasible* post-FMSC confidence interval constructed directly from  $Q_{FMSC}$ . This interval has asymptotic coverage probability *exactly* equal to its nominal level as it correctly accounts for the effect of moment selection on the asymptotic distribution of the estimators. Unfortunately it cannot be used in practice because it requires knowledge of  $\tau$ , for which no consistent estimator exists. As such, this interval serves as a benchmark against which to judge various feasible procedures that do not require knowledge of  $\tau$ . For certain parameter values this interval is shorter than the valid interval but the improvement is not uniform and indeed cannot be. Just as the FMSC itself cannot provide a uniform reduction in AMSE relative to the valid estimator, the infeasible post-FMSC cannot provide a corresponding reduction in width. In both cases, however, improvements are possible when  $\tau$  is expected to be small, the setting in which this paper assumes that an applied researcher finds herself. The potential reductions in width can be particularly dramatic for larger values of  $\alpha$ . The question remains: is there any way to capture these gains using a *feasible* procedure?

Now, consider the 2-Step confidence interval procedure from Algorithm 4.3. We can implement an equivalent procedure without simulation as follows. First we construct a  $(1 - \alpha_1) \times 100\%$  confidence interval for  $\hat{\tau}$  using  $T = \sigma Z_1 + \tau$  where  $Z_1$  is standard normal. Next we construct a  $(1 - \alpha_2) \times 100\%$  based on  $Q_{FMSC}$  for each  $\tau^*$  in this interval. Finally we take the upper and lower bounds over all of the resulting intervals. This interval is guaranteed to have asymptotic coverage probability of at least  $1 - (\alpha_1 + \alpha_2)$  by an argument essentially identical to the proof of Theorem 4.4. Protection against under-coverage, however, comes at the expense of extreme conservatism, particularly for larger values of  $\alpha$ . Numerical values for the coverage and median expected width of this interval appear in online appendix F.2. From both the numerical calculations and the theoretical result given in Theorem 4.4 we see that the 2-Step systematically *over-covers* and hence *cannot* produce an interval shorter than the textbook CI for the valid estimator.

Now consider the 1-Step confidence interval from Algorithm 4.2. Rather than first constructing a confidence region for  $\tau$  and then taking upper and lower bounds, 1-Step interval simply takes  $\hat{\tau}$  in place of  $\tau$  and then constructs a confidence interval from  $Q_{FMSC}$  exactly as in the infeasible interval described above.<sup>24</sup> Unlike its 2-Step counterpart, this interval comes with no generic theoretical guarantees, so I use the characterization from above to directly calculate its asymptotic coverage and expected relative width. The results appear in Tables 4 and 5. The 1-Step interval effectively

<sup>22</sup> I provide code to evaluate both  $F_{FMSC}$  and  $Q_{FMSC}$  in my R package `fmscr`, available at <https://github.com/fditraglia/fmscr>.

<sup>23</sup> The population first-stage  $R$ -squared with only  $z$  in the instrument set is  $1/9$ .

<sup>24</sup> As in the construction of the naïve interval, I take the shortest possible interval based on  $Q_{FMSC}$  rather than an equal-tailed interval. Additional results for an equal-tailed version of this one-step procedure are available upon request. Their performance is similar.

**Table 1**Asymptotic coverage probability of naïve  $(1 - \alpha) \times 100\%$  confidence interval.

(a) OLS versus TSLS							(b) Choosing IVs								
$\alpha = 0.05$	$\tau$						$\alpha = 0.05$	$\tau$							
	0	1	2	3	4	5		0	1	2	3	4	5		
$\pi^2$	0.1	91	81	57	41	45	58	$\gamma^2$	0.1	93	89	84	85	91	94
	0.2	91	83	63	58	70	84		0.2	92	87	76	74	83	91
	0.3	92	84	69	73	86	93		0.3	92	85	71	65	74	86
	0.4	92	85	76	84	93	95		0.4	91	85	68	59	67	80
$\alpha = 0.1$	$\tau$						$\alpha = 0.1$	$\tau$							
	0	1	2	3	4	5		0	1	2	3	4	5		
$\pi^2$	0.1	83	70	45	35	42	55	$\gamma^2$	0.1	87	82	76	79	86	89
	0.2	84	72	53	52	67	81		0.2	85	78	66	67	79	87
	0.3	85	74	60	68	83	89		0.3	84	76	61	59	71	82
	0.4	86	76	68	80	89	90		0.4	84	75	57	52	63	77
$\alpha = 0.2$	$\tau$						$\alpha = 0.2$	$\tau$							
	0	1	2	3	4	5		0	1	2	3	4	5		
$\pi^2$	0.1	70	54	31	27	37	50	$\gamma^2$	0.1	75	69	64	70	77	80
	0.2	71	57	39	45	62	74		0.2	73	64	53	59	71	78
	0.3	73	59	49	61	75	79		0.3	72	62	47	50	64	75
	0.4	74	62	58	72	79	80		0.4	72	60	43	44	58	71

**Table 2**

Asymptotic expected width of naïve confidence interval relative to that of the valid estimator. Values are given in percentage points.

(a) OLS versus TSLS							(b) Choosing IVs								
		$\tau$							$\tau$						
		0	1	2	3	4	5			0	1	2	3	4	5
$\pi^2$	0.1	42	44	48	55	64	73	$\gamma^2$	0.1	77	80	87	94	98	100
	0.2	53	56	64	74	85	92		0.2	66	69	77	86	93	98
	0.3	62	66	76	87	95	99		0.3	60	62	69	79	88	94
	0.4	69	74	85	94	99	100		0.4	55	57	64	73	83	90

**Table 3**Width of shortest possible  $(1 - \alpha) \times 100\%$  post-FMSC confidence interval constructed directly from  $Q_{FMSC}$  using knowledge of  $\tau$ . Values are given in percentage points.

(a) OLS versus TSLS							(b) Choosing IVs								
$\alpha = 0.05$	$\tau$						$\alpha = 0.05$	$\tau$							
	0	1	2	3	4	5		0	1	2	3	4	5		
$\pi^2$	0.1	99	92	85	89	95	101	$\gamma^2$	0.1	92	97	106	111	109	102
	0.2	97	91	94	102	110	117		0.2	93	94	101	109	115	114
	0.3	94	94	102	111	117	109		0.3	95	93	97	105	112	117
	0.4	92	97	107	114	107	100		0.4	97	92	94	101	108	115
$\alpha = 0.1$	$\tau$						$\alpha = 0.1$	$\tau$							
	0	1	2	3	4	5		0	1	2	3	4	5		
$\pi^2$	0.1	88	81	85	91	99	107	$\gamma^2$	0.1	89	97	108	113	108	101
	0.2	89	88	97	107	116	123		0.2	86	93	104	113	118	109
	0.3	86	93	105	115	119	103		0.3	86	90	100	109	117	121
	0.4	87	98	111	116	104	100		0.4	88	88	96	105	114	121
$\alpha = 0.2$	$\tau$						$\alpha = 0.2$	$\tau$							
	0	1	2	3	4	5		0	1	2	3	4	5		
$\pi^2$	0.1	48	55	84	96	106	116	$\gamma^2$	0.1	86	96	111	115	105	101
	0.2	65	80	101	114	125	117		0.2	78	89	108	119	118	104
	0.3	74	90	111	123	112	101		0.3	72	84	103	116	125	112
	0.4	80	97	116	115	102	100		0.4	67	79	99	112	123	128

“splits the difference” between the two-step interval and the naïve procedure. While it can under-cover, the size distortions are quite small, particularly for  $\alpha = 0.1$  and  $0.05$ . At the same time, when  $\tau$  is relatively small this procedure can yield shorter intervals. While a full investigation of this phenomenon is beyond the scope of the present paper, these calculations suggest a plausible way forward for post-FMSC inference that is less conservative than the two-step procedure from Algorithm 4.3 by directly calculating the relevant quantities from the limit distribution of interest. This is possible because  $\pi$  and  $\gamma^2$  are both consistently estimable. And for any par-

ticular value of these parameters, the worst-case value of  $\tau$  is *interior*. Using this idea, one could imagine specifying a maximum allowable size distortion and then designing a confidence interval to minimize width, possibly incorporating some prior restriction on the likely magnitude of  $\tau$ . Just as the FMSC aims to achieve a favorable trade-off between bias and variance, such a confidence interval procedure could aim to achieve a favorable trade-off between width and coverage. It would also be interesting to pursue analogous calculations for the minimum AMSE averaging estimator from Section 4.2.

**Table 4**

OLS vs TSLS example: shortest 1-Step CI.

(a) Coverage probability							(b) Relative width							
$\alpha = 0.05$							$\alpha = 0.05$							
	$\tau$							$\tau$						
	0	1	2	3	4	5		0	1	2	3	4	5	
$\pi^2$	0.1	93	94	95	94	91	90	0.1	93	93	95	97	99	102
	0.2	95	95	95	93	91	91	0.2	96	97	99	104	106	109
	0.3	95	96	94	92	92	94	0.3	97	99	102	106	108	107
	0.4	96	95	94	93	95	95	0.4	98	100	105	108	106	103
$\alpha = 0.1$							$\alpha = 0.1$							
	$\tau$							$\tau$						
	0	1	2	3	4	5		0	1	2	3	4	5	
$\pi^2$	0.1	89	89	88	86	82	80	0.1	90	91	92	97	99	102
	0.2	91	91	88	85	83	85	0.2	94	96	100	105	108	110
	0.3	92	91	87	85	87	90	0.3	96	100	104	108	109	106
	0.4	92	90	87	87	90	91	0.4	97	101	106	108	106	103
$\alpha = 0.2$							$\alpha = 0.2$							
	$\tau$							$\tau$						
	0	1	2	3	4	5		0	1	2	3	4	5	
$\pi^2$	0.1	84	80	71	67	65	64	0.1	83	84	87	93	99	103
	0.2	85	80	71	70	70	76	0.2	91	92	96	105	109	110
	0.3	84	79	73	72	78	81	0.3	93	97	104	109	108	106
	0.4	84	79	74	77	81	81	0.4	95	100	107	108	105	102

**Table 5**

Choosing IVs example: shortest 1-Step CI.

(a) Coverage probability							(b) Relative width							
$\alpha = 0.05$							$\alpha = 0.05$							
	$\tau$							$\tau$						
	0	1	2	3	4	5		0	1	2	3	4	5	
$\gamma^2$	0.1	96	95	94	93	94	95	0.1	98	100	104	106	106	104
	0.2	96	96	95	93	93	94	0.2	97	99	103	106	108	108
	0.3	95	95	95	93	92	92	0.3	97	98	101	104	107	109
	0.4	95	95	95	94	92	91	0.4	97	97	99	103	106	108
$\alpha = 0.1$							$\alpha = 0.1$							
	$\tau$							$\tau$						
	0	1	2	3	4	5		0	1	2	3	4	5	
$\gamma^2$	0.1	92	90	88	88	89	91	0.1	98	100	104	107	106	104
	0.2	92	91	88	86	86	89	0.2	97	98	103	107	109	107
	0.3	92	91	89	86	85	87	0.3	96	97	101	105	108	109
	0.4	91	91	89	86	84	84	0.4	95	96	100	103	107	109
$\alpha = 0.2$							$\alpha = 0.2$							
	$\tau$							$\tau$						
	0	1	2	3	4	5		0	1	2	3	4	5	
$\gamma^2$	0.1	83	80	76	77	80	81	0.1	98	100	105	107	106	103
	0.2	84	80	75	73	76	80	0.2	94	97	104	108	109	107
	0.3	85	81	75	71	72	77	0.3	93	96	101	106	109	109
	0.4	84	80	73	72	69	74	0.4	89	93	97	105	108	110

## 5. Simulation results

### 5.1. OLS versus TSLS example

I begin by examining the performance of the FMSC and averaging estimator in the OLS versus TSLS example. All calculations in this section are based on the formulas from Sections 3.2 and 4.2 with 10,000 simulation replications. The data generating process is given by

$$y_i = 0.5x_i + \epsilon_i \quad (21)$$

$$x_i = \pi(z_{1i} + z_{2i} + z_{3i}) + v_i \quad (22)$$

with  $(\epsilon_i, v_i, z_{1i}, z_{2i}, z_{3i}) \sim \text{iid } N(0, \delta)$

$$\delta = \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix}, \quad \delta_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 - \pi^2 \end{bmatrix}, \quad \delta_2 = I_3/3 \quad (23)$$

for  $i = 1, \dots, N$  where  $N$ ,  $\rho$  and  $\pi$  vary over a grid. The goal is to estimate the effect of  $x$  on  $y$ , in this case 0.5, with minimum MSE either by choosing between OLS and TSLS estimators or by averaging them. To ensure that the finite-sample MSE of the TSLS estimator exists, this DGP includes three instruments leading

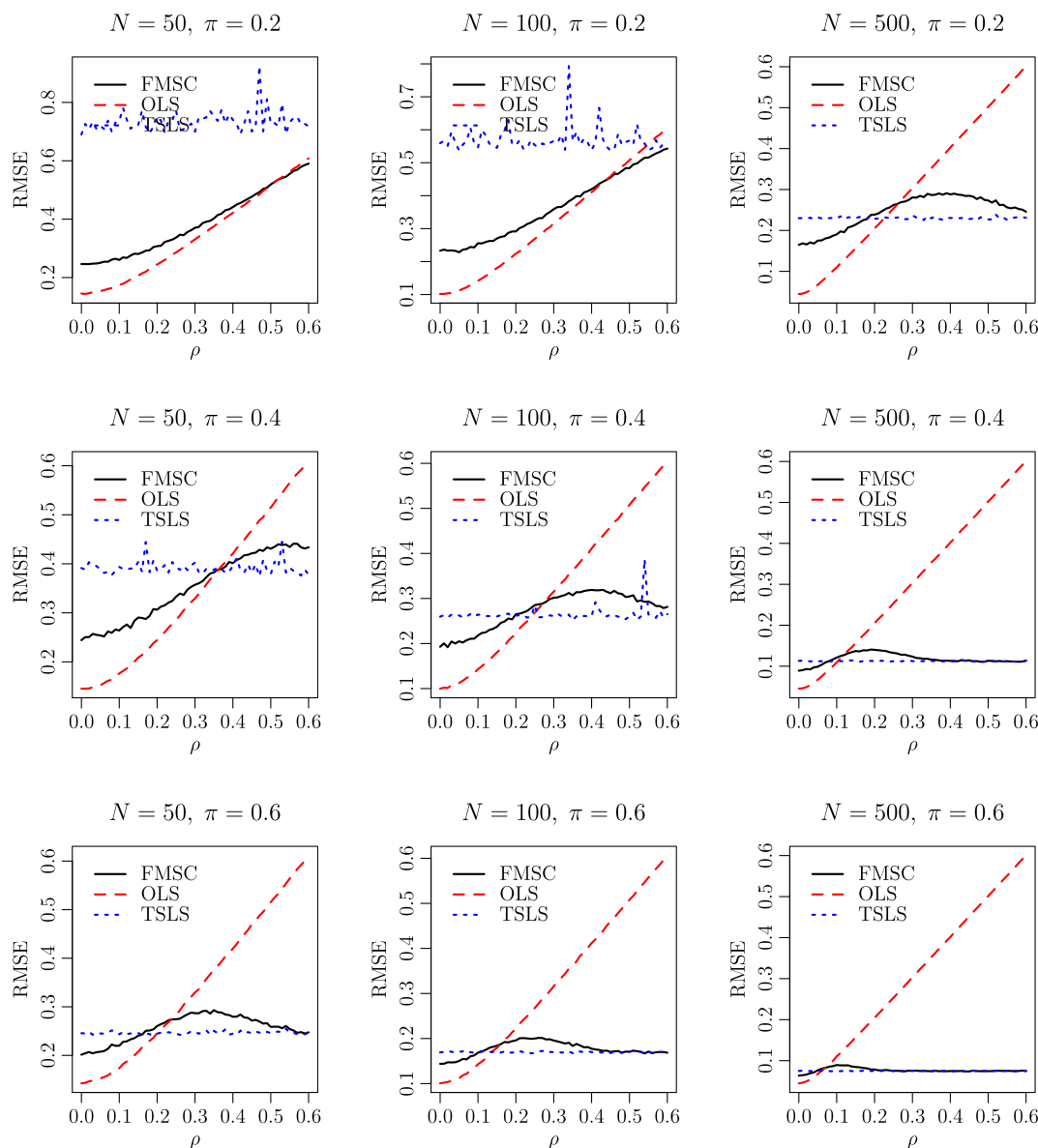
to two overidentifying restrictions (Phillips, 1980).<sup>25</sup> This design satisfies regularity conditions that are sufficient for Theorem 3.2 – in particular it satisfies Assumption E.1 from online appendix E – and keeps the variance of  $x$  fixed at one so that  $\pi = \text{Cor}(x_i, z_{1i} + z_{2i} + z_{3i})$  and  $\rho = \text{Cor}(x_i, \epsilon_i)$ . 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.

Fig. 1 compares the root mean-squared error (RMSE) of the post-FMSC estimator to those of the OLS and TSLS estimators.<sup>26</sup> 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. In practice, of course,  $\rho$  is unknown so we cannot tell which of OLS and TSLS is to be preferred

<sup>25</sup> Alternatively, one could use fewer instruments in the DGP and work with trimmed MSE, as described in online appendix C.

<sup>26</sup> Note that, while the first two moments of the TSLS estimator exist in this simulation design, none of its higher moments do. This can be seen from the simulation results: even with 10,000 replications, the RMSE of the TSLS estimator shows a noticeable degree of simulation error.





**Fig. 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 Eqs. (22)–(23) using the formulas described in Section 3.2.

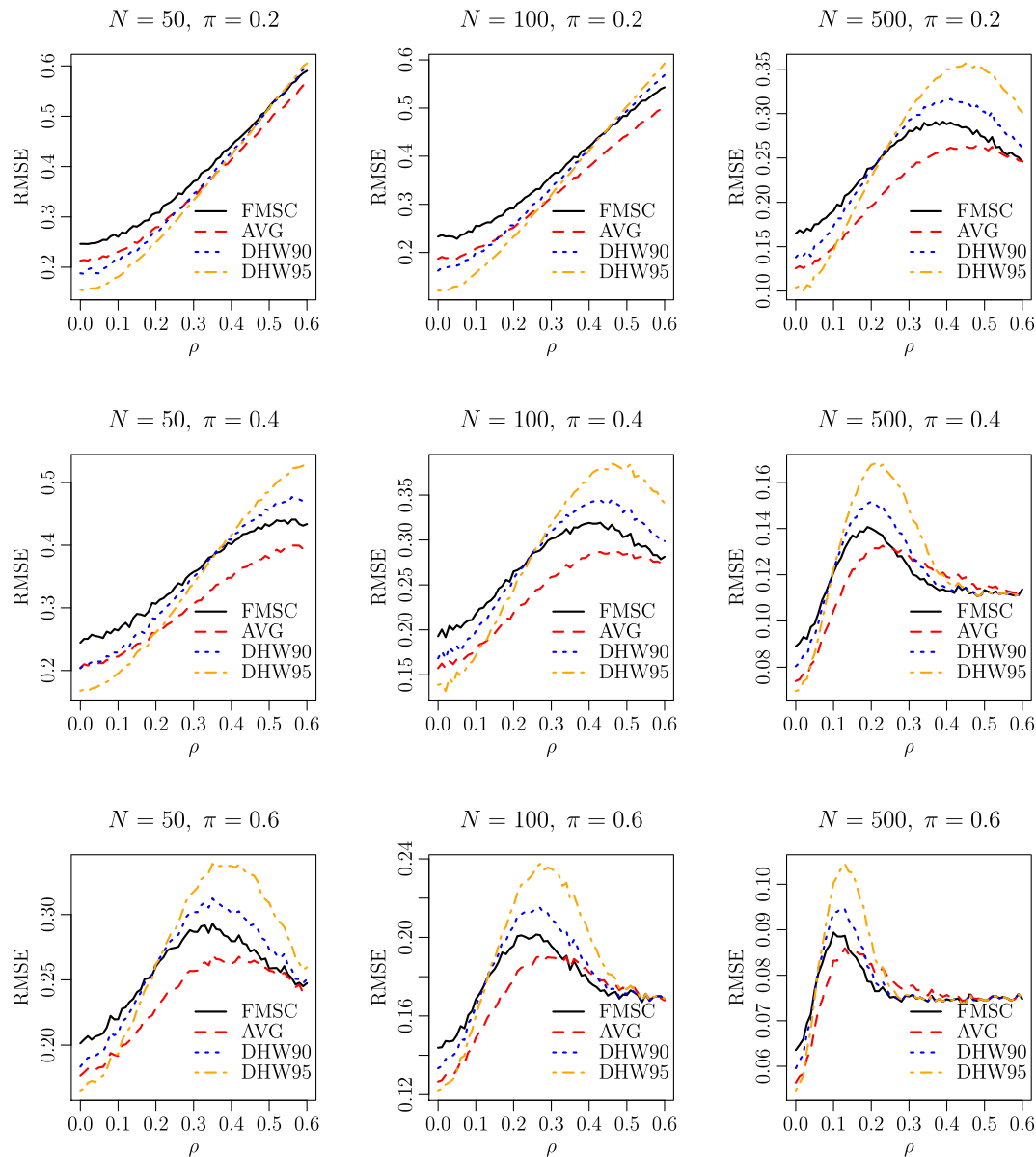
*a priori*. If we make it our policy to always use TSLS we protect ourselves against bias at the potential cost of very high variance. If, on the other hand, we make it our policy to always use OLS then we protect ourselves against high variance at the potential cost of severe bias. The FMSC represents a compromise between these two extremes that does not require advance knowledge of  $\rho$ . 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>27</sup> For this reason it does not attain an RMSE equal to the lower envelope of the OLS and TSLS estimators. The larger the RMSE difference between OLS and TSLS, however, the closer the FMSC comes to this lower envelope: costly mistakes are rare.

As shown above, the FMSC takes a very special form in this example: it is equivalent to a DHW test with  $\alpha \approx 0.16$ . Accordingly, Fig. 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 Fig. 1 that OLS has a lower RMSE than TSLS. 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. In addition to comparing the FMSC to DHW pre-test estimators, Fig. 2 also presents the finite-sample RMSE of the minimum-AMSE moment average estimator presented in Eqs. (12) and (13). The performance of the moment average estimator is very strong: it provides the lowest worst-case RMSE and improves uniformly on the FMSC for all but the largest values of  $\rho$ .

Because this example involves a scalar target parameter, no selection or averaging scheme can provide a *uniform* improvement over the minimax estimator, namely TSLS. But the cost of

<sup>27</sup> For more discussion of this point, see Section 4.



**Fig. 2.** 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 Eqs. (22)–(23) using the formulas described in Sections 3.2 and 4.2.

protection against the worst case is extremely poor performance when  $\pi$  and  $N$  are small. When this is the case, there is a strong argument for preferring the FMSC or minimum-AMSE estimator: we can reap the benefits of OLS when  $\rho$  is small without risking the extremely large biases that could result if  $\rho$  is in fact large.

Further simulation results for  $\pi \in \{0.01, 0.05, 0.1\}$  appear in online appendix G.4. For these parameter values the TSLS estimator suffers from a weak instrument problem leading the FMSC to substantially outperform the TSLS estimator. See online appendix G.4 for a more detailed discussion.

## 5.2. Choosing instrumental variables example

I now evaluate the performance of FMSC in the instrument selection example described in Section 3.3 using the following simulation design:

$$y_i = 0.5x_i + \epsilon_i \quad (24)$$

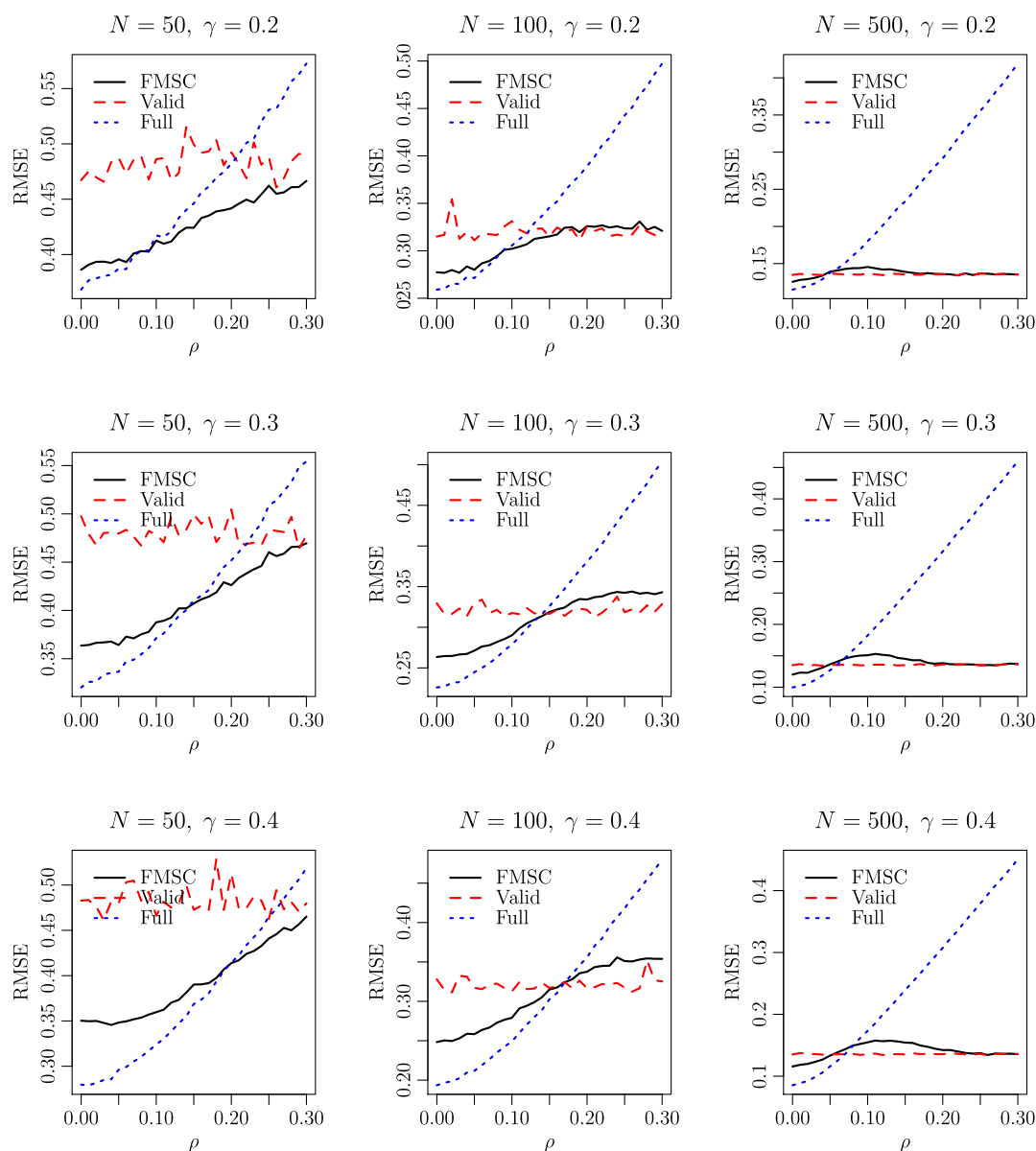
$$x_i = (z_{1i} + z_{2i} + z_{3i})/3 + \gamma w_i + v_i \quad (25)$$

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

$$\mathcal{V} = \begin{bmatrix} v_1 & 0 \\ 0 & v_2 \end{bmatrix}, \quad v_1 = \begin{bmatrix} 1 & (0.5 - \gamma\rho) & \rho \\ (0.5 - \gamma\rho) & (8/9 - \gamma^2) & 0 \\ \rho & 0 & 1 \end{bmatrix}, \quad (26)$$

and  $v_2 = I_3/3$

This setup keeps the variance of  $x$  fixed at one and the endogeneity of  $x$ ,  $\text{Cor}(x, \epsilon)$ , fixed at 0.5 while allowing the relevance,  $\gamma = \text{Cor}(x, w)$ , and endogeneity,  $\rho = \text{Cor}(w, \epsilon)$ , 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. This design satisfies the sufficient conditions for Theorem 3.5 given in



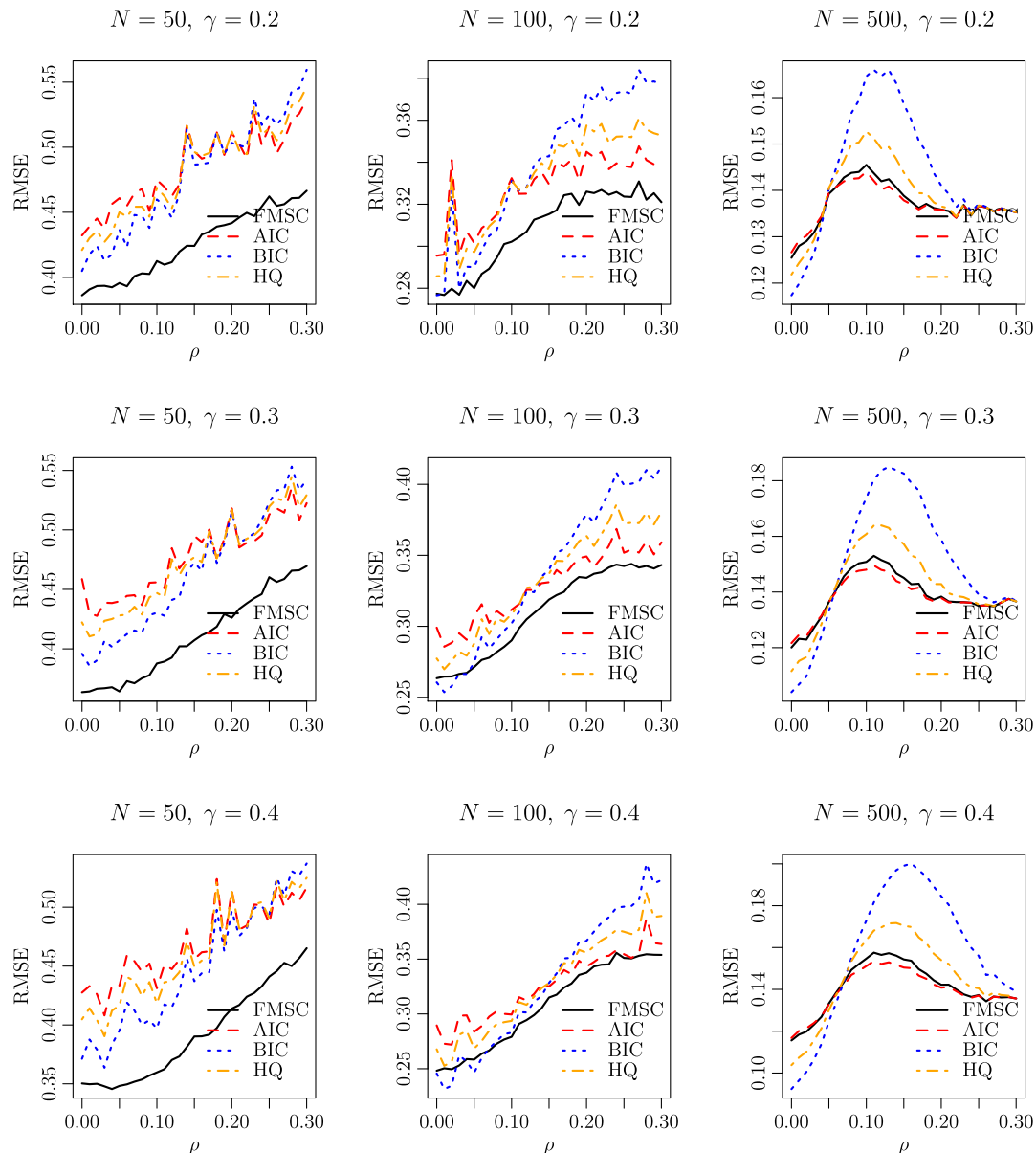
**Fig. 3.** RMSE values for the valid estimator, including only  $(z_1, z_2, z_3)$ , the full estimator, including  $(z_1, z_2, z_3, w)$ , and the post-Focused Moment Selection Criterion (FMSC) estimator based on 20,000 simulation draws from the DGP given in Eqs. (25)–(26) using the formulas described in Section 3.3.

Assumption E.2 from online appendix E. When  $\gamma = 0$ , it is a special case of the DGP from Section 5.1.

As in Section 5.1, 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>28</sup> All estimators in this section are calculated via TSLS without a constant term using the expressions from Section 3.3 and 20,000 simulation replications.

Fig. 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 Fig. 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. 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

<sup>28</sup> Alternatively, one could use fewer instruments for the valid estimator and compare the results using *trimmed* MSE. For details, see online appendix C.



**Fig. 4.** 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 Eqs. (25)–(26) using the formulas described in Section 3.3.

smaller sample sizes. The FMSC never performs much worse than the valid estimator, however, and often performs substantially better, particularly for small sample sizes.

I now compare the FMSC to the GMM moment selection criteria of Andrews (1999), which take the form  $MSC(S) = J_n(S) - h(|S|)\kappa_n$ , where  $J_n(S)$  is the  $J$ -test statistic under moment set  $S$  and  $-h(|S|)\kappa_n$  is a “bonus term” that 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. 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. Fig. 4 gives the RMSE values for the post-FMSC estimator alongside those of the post-GMM-BIC, HQ and AIC estimators. I calculate the  $J$ -test statistic using a centered

covariance matrix estimator, following the recommendation of Andrews (1999). For small sample sizes, the GMM-BIC, AIC and HQ are quite erratic: indeed for  $N = 50$  the FMSC has a uniformly smaller RMSE. This problem comes from the fact that the  $J$ -test statistic can be very badly behaved in small samples.<sup>29</sup> 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. For a discussion of a combined strategy based on the GMM information criteria of Andrews (1999) and the canonical correlations information criteria of Hall and Peixe (2003), see online appendix G.2. For a comparison with the downward  $J$ -test, see online appendix G.1. online appendix G.4 presents results for a modified simulation

<sup>29</sup> For more details, see online appendix G.1.



**Table 6**Naïve CI, OLS vs IV example,  $N = 100$ .

(a) Coverage probability							(b) Average relative width							
$\alpha = 0.05$	$\rho$						$\alpha = 0.05$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\pi^2$	0.1	92	83	56	36	38	55	0.1	39	37	31	50	56	69
	0.2	91	84	63	56	73	85	0.2	52	54	64	74	86	95
	0.3	92	85	70	73	88	93	0.3	60	65	75	87	96	100
	0.4	93	85	76	83	93	94	0.4	68	73	84	95	100	100
$\alpha = 0.1$	$\rho$						$\alpha = 0.1$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\pi^2$	0.1	87	73	42	33	41	54	0.1	38	39	43	49	54	69
	0.2	85	76	50	53	73	83	0.2	52	55	61	74	87	95
	0.3	86	75	63	71	86	88	0.3	61	65	76	88	97	100
	0.4	86	74	68	83	89	89	0.4	69	74	85	95	100	100
$\alpha = 0.2$	$\rho$						$\alpha = 0.2$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\pi^2$	0.1	72	57	29	26	38	53	0.1	40	41	40	49	57	68
	0.2	74	59	40	45	65	77	0.2	52	55	63	73	86	95
	0.3	74	59	46	62	78	79	0.3	61	65	74	88	96	99
	0.4	75	61	59	74	81	78	0.4	69	73	85	95	100	100

**Table 7**Naïve CI, choosing IVs example,  $N = 100$ .

(a) Coverage probability							(b) Average relative width							
$\alpha = 0.05$	$\rho$						$\alpha = 0.05$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\gamma^2$	0.1	90	80	77	83	88	93	0.1	72	72	80	91	97	99
	0.2	89	78	66	69	82	92	0.2	62	61	72	83	91	98
	0.3	87	78	59	60	77	86	0.3	56	54	58	71	87	94
	0.4	89	77	55	50	67	81	0.4	49	47	53	67	80	87
$\alpha = 0.1$	$\rho$						$\alpha = 0.1$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\gamma^2$	0.1	84	72	73	77	86	89	0.1	71	67	81	91	99	98
	0.2	83	72	58	66	78	86	0.2	60	61	67	78	89	92
	0.3	84	70	51	55	73	82	0.3	55	54	59	72	82	94
	0.4	81	65	46	48	63	78	0.4	51	47	55	68	78	89
$\alpha = 0.2$	$\rho$						$\alpha = 0.2$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\gamma^2$	0.1	76	63	60	70	77	79	0.1	71	71	79	92	96	98
	0.2	75	59	49	58	73	77	0.2	60	60	68	81	91	97
	0.3	69	54	39	48	67	76	0.3	55	56	61	72	84	95
	0.4	71	53	34	40	57	74	0.4	51	48	53	64	78	89

experiment in which the valid estimator suffers from a weak instrument problem. The FMSC performs very well in this case.

### 5.3. Confidence interval simulations

I now revisit the simulation experiments introduced above in Sections 5.1 and 5.2 to evaluate the finite-sample performance of the confidence intervals whose asymptotic performance was studied in Section 4.4. All results in this section are based on 1000 simulation replications from the relevant DGP. Coverage probabilities and relative widths are all given in percentage points, rounded to the nearest whole percent. In the interest of brevity I present only results for  $N = 100$ . Likewise, for the two-step confidence intervals I present results only for  $\alpha_1 = \alpha/4$ ,  $\alpha_2 = 3\alpha/4$ . Simulation results for  $N = 50$  and 500 and other configurations of  $\alpha_1, \alpha_2$  are available upon request. Taking  $N = 100$  has the advantage of making the tables in this section directly comparable to those of Section 4.4. Because I set  $\sigma_x^2 = \sigma_\epsilon^2 = 1$  in both simulation experiments, this implies that  $\sqrt{N}\rho = \sqrt{N}\text{Cor}(x_i, \epsilon_i) = \tau$  in the OLS versus TSLS example and  $\sqrt{N}\rho = \sqrt{N}\text{Cor}(w_i, \epsilon_i) = \tau$  in the choosing

IVs example. Thus when  $N = 100$ , taking  $\rho \in \{0, 0.1, \dots, 0.5\}$  is the finite-sample analogue of  $\tau \in \{0, 1, \dots, 5\}$ .

To begin, Tables 6 and 7 present the coverage probability and average relative width of a naïve confidence interval that ignores the effects of moment selection on inference. These are the finite-sample analogues of Tables 1 and 2. For the OLS versus IV example, expected relative width is calculated relative to a textbook confidence interval for the TSLS estimator while for the choosing IVs example it is calculated relative to a textbook confidence interval for the valid estimator that excludes  $w$  from the instrument set. As in the asymptotic calculations presented above, we find that the naïve procedure suffers from severe size distortions but results in much shorter intervals. Results for the 2-Step confidence interval appear in online appendix G.3. With a small allowance for sampling variability, we see that the 2-step intervals indeed provide uniform coverage no lower than their nominal level but result in wider intervals than simply using TSLS or the valid estimator, respectively. Tables 8 and 9, the finite-sample analogues of Tables 4 and 5, present results for the one-step confidence interval that assumes  $\hat{\tau} = \tau$ . As expected from the asymptotic calculations from Section 4.4, this interval presents

**Table 8**1-Step shortest CI, OLS vs IV example,  $N = 100$ .

(a) Coverage probability							(b) Average relative width							
$\alpha = 0.05$	$\rho$						$\alpha = 0.05$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\pi^2$	0.1	95	97	97	93	80	72	0.1	92	91	94	94	95	99
	0.2	95	98	94	88	84	89	0.2	96	96	99	102	107	112
	0.3	96	96	95	88	91	96	0.3	97	98	103	107	111	110
	0.4	96	96	92	90	96	95	0.4	98	100	105	109	108	102
$\alpha = 0.1$	$\rho$						$\alpha = 0.1$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\pi^2$	0.1	94	93	91	82	73	68	0.1	88	89	91	93	95	102
	0.2	91	92	87	79	81	88	0.2	94	96	99	105	110	114
	0.3	93	91	86	82	90	93	0.3	95	98	104	110	111	107
	0.4	91	90	85	89	92	90	0.4	97	101	106	110	106	101
$\alpha = 0.2$	$\rho$						$\alpha = 0.2$	$\rho$						
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5	
$\pi^2$	0.1	83	79	72	64	61	63	0.1	81	83	81	91	97	105
	0.2	86	81	70	67	75	83	0.2	89	91	98	106	113	114
	0.3	84	80	73	75	84	82	0.3	93	96	104	111	111	105
	0.4	83	78	75	81	84	78	0.4	96	99	108	110	105	101

**Table 9**1-Step shortest CI, choosing IVs example,  $N = 100$ .

(a) Coverage probability							(b) Average relative width								
$\alpha = 0.05$	$\rho$						$\alpha = 0.05$	$\rho$							
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5		
$\gamma^2$	0.1	93	90	89	89	92	94	$\gamma^2$	0.1	97	99	102	106	107	106
	0.2	93	92	89	85	90	95		0.2	97	97	101	104	108	109
	0.3	92	93	88	86	88	92		0.3	96	95	97	101	106	110
	0.4	94	94	90	83	83	87		0.4	94	94	96	99	104	108
$\alpha = 0.1$	$\rho$						$\alpha = 0.1$	$\rho$							
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5		
$\gamma^2$	0.1	89	84	84	85	90	91	$\gamma^2$	0.1	97	96	104	108	108	105
	0.2	90	89	82	81	86	90		0.2	95	96	100	105	109	107
	0.3	90	89	83	78	82	88		0.3	94	94	98	103	107	111
	0.4	87	89	82	78	77	85		0.4	93	92	96	102	106	110
$\alpha = 0.2$	$\rho$						$\alpha = 0.2$	$\rho$							
	0	0.1	0.2	0.3	0.4	0.5		0	0.1	0.2	0.3	0.4	0.5		
$\gamma^2$	0.1	81	76	73	76	80	80	$\gamma^2$	0.1	95	98	104	110	109	104
	0.2	83	77	74	75	80	80		0.2	90	95	102	110	113	109
	0.3	80	74	69	69	78	82		0.3	90	93	99	105	111	113
	0.4	81	76	68	67	70	80		0.4	87	89	95	103	109	113

a good trade-off between the naïve and 2-step CIs: it can yield shorter intervals with far smaller size distortions. Because it is also simple to compute, the 1-Step interval could prove quite valuable in practice. It would be interesting to explore this interval further both theoretically and in simulation studies.

## 6. Empirical example: Geography or institutions?

Carstensen and Gundlach (2006) address a controversial question from the development literature: what is the causal effect of geography on income per capita after controlling for the quality of institutions? A number of well-known studies find little or no direct effect of geographic endowments (Acemoglu et al., 2001; Rodrik et al., 2004; Easterly and Levine, 2003). 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) extend Sachs's

work using the following baseline regression for a sample of 44 countries:

$$\ln gdp_i = \beta_1 + \beta_2 \cdot institutions_i + \beta_3 \cdot malaria_i + \epsilon_i. \quad (27)$$

This model augments the baseline specification of Acemoglu et al. (2001) to include a direct effect of malaria transmission which, like institutions, is treated as endogenous.<sup>30</sup> Considering a variety of measures of both institutions and malaria transmission, and a number of instrument sets, Carstensen and Gundlach (2006) find large negative effects of malaria transmission, lending support to Sachs's conclusion.

In this section, I revisit and expand upon the instrument selection exercise given in Table 2 of Carstensen and Gundlach (2006) using the FMSC and corrected confidence intervals described above. All results in this section are calculated by TSLS using the formulas from Section 3.3 and the variables described in Table 10, with  $\ln gdp_i$  as the outcome variable and  $rule$  and  $malfal$  as

<sup>30</sup> Due to a lack of data for certain instruments, Carstensen and Gundlach (2006) work with a smaller sample of countries than Acemoglu et al. (2001).

**Table 10**  
Description of variables for Empirical Example.

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

measures of institutions and malaria transmission. In this exercise I imagine two hypothetical econometricians. The first, like [Sachs \(2003\)](#) and [Carstensen and Gundlach \(2006\)](#), seeks the best possible estimate of the causal effect of malaria transmission,  $\beta_3$ , after controlling for institutions by selecting over a number of possible instruments. The second, in contrast, seeks the best possible estimate of the causal effect of institutions,  $\beta_2$ , after controlling for malaria transmission by selecting over the same collection of instruments. After estimating their desired target parameters, both econometricians also wish to report valid confidence intervals that account for the additional uncertainty introduced by instrument selection. For the purposes of this example, to illustrate the results relevant to each of my hypothetical researchers, I take each of  $\beta_2$  and  $\beta_3$  in turn as the target parameter.<sup>31</sup>

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 *lnmort* 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 *lnmort*, *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 10](#). [Table 11](#) lists the eight instrument sets considered here, along with TSLS estimates and traditional 95% confidence intervals for each.<sup>32</sup>

[Table 12](#) presents FMSC and “positive-part” FMSC results for instrument sets 1–8. The positive-part FMSC sets a negative squared bias estimate to zero when estimating AMSE. If the squared bias estimate is positive, FMSC and positive-part FMSC coincide; if 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 5.2](#), available upon request, reveal that the positive-part FMSC never performs worse than the ordinary FMSC and

sometimes performs slightly better, suggesting that it may be preferable in real-world applications. For each criterion the table 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). 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.

[Table 13](#) presents three alternative post-selection confidence intervals for each of the instrument selection exercises from [Table 12](#): Naïve, 1-Step, and 2-Step. These intervals are constructed as described in the simulation experiments from [Section 5.3](#). The simulation-based intervals are based on 10,000 random draws. For the two-step interval I take  $\alpha = \delta = 0.025$  which guarantees asymptotic coverage of at least 95%. From the resulting intervals, we see that both of our two hypothetical econometricians would report a statistically significant result even after accounting for the effects of instrument selection on inference and in spite of the conservatism of the 2-Step interval.

## 7. Conclusion

This paper has introduced the FMSC, a proposal to choose moment conditions using AMSE. The criterion performs well in simulations, and the framework under which it is derived can be used produce confidence intervals that adjust for the effects of moment selection on subsequent inference. Moment selection is not a panacea, but the FMSC and related confidence interval procedures can yield sizeable benefits in empirically relevant settings. 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.

## 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 3.2, 3.5.** The proofs of both results are similar and standard, so I provide only a sketch of the argument for [Theorem 3.5](#). First substitute the DGP into the expression for  $\hat{\beta}_S$  and rearrange so that the left-hand side becomes  $\sqrt{n}(\hat{\beta}_S - \beta)$ . 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 CLT.  $\square$

<sup>31</sup> A researcher interested in both  $\beta_2$  and  $\beta_3$ , however, should not proceed in this fashion, as it could lead to contradictory inferences. Instead, she should define a target parameter that includes both  $\beta_2$  and  $\beta_3$ .

<sup>32</sup> The results for the baseline instrument presented in panel 1 of [Table 11](#) are slightly different from those in [Carstensen and Gundlach \(2006\)](#) as I exclude Vietnam to keep the sample fixed across instrument sets.

**Table 11**

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

	1		2		3		4	
	rule	malfal	rule	malfal	rule	malfal	rule	malfal
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 Openness		Baseline Europe	
	5		6		7		8	
	rule	malfal	rule	malfal	rule	malfal	rule	malfal
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 Climate Europe		Baseline Climate Openness		Baseline Openness Europe		Baseline Climate Openness Europe	

**Table 12**

FMSC and positive-part FMSC values corresponding to the instrument sets from Table 11.

	$\mu = \text{malfal}$			$\mu = \text{rule}$		
	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 13**

Post-selection CIs for the instrument selection exercise from Table 12.

	$\mu = \text{malfal}$		$\mu = \text{rule}$	
	FMSC	posFMSC	FMSC	posFMSC
Naïve	(−1.66, −0.50)	(−1.66, −0.50)	(0.53, 1.14)	(0.59, 1.27)
1-Step	(−1.58, −0.61)	(−1.57, −0.62)	(0.53, 1.12)	(0.64, 1.21)
2-Step	(−1.69, −0.48)	(−1.69, −0.48)	(0.45, 1.22)	(0.54, 1.31)

**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 Theorem 3.4.** By Theorem 3.3,  $\sqrt{n}(\hat{\beta}_{OLS} - \tilde{\beta}_{TSLS}) \rightarrow_d N(\tau/\sigma_x^2, \Sigma)$  where  $\Sigma = \sigma_\epsilon^2(1/\gamma^2 - 1/\sigma_x^2)$ . Thus, under  $H_0: \tau = 0$ , the DHW test statistic

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

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

$$\begin{aligned}\hat{V} &= \hat{\sigma}_\epsilon^2\hat{\sigma}_x^2\left(\frac{\hat{\sigma}_v^2}{\hat{\gamma}^2}\right) = \hat{\sigma}_\epsilon^2\hat{\sigma}_x^2\left(\frac{\hat{\sigma}_x^2 - \hat{\gamma}^2}{\hat{\gamma}^2}\right) \\ &= \hat{\sigma}_\epsilon^2\hat{\sigma}_x^4\left(\frac{1}{\hat{\gamma}^2} - \frac{1}{\hat{\sigma}_x^2}\right) = \hat{\sigma}_x^4\hat{\Sigma}\end{aligned}$$

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

$$\begin{aligned}\hat{\tau}^2 &= [n^{-1/2}\mathbf{x}'(\mathbf{y} - \mathbf{x}\tilde{\beta})]^2 = n^{-1}[\mathbf{x}'\mathbf{x}(\hat{\beta} - \tilde{\beta})]^2 \\ &= n^{-1}[n\hat{\sigma}_x^2(\hat{\beta} - \tilde{\beta})]^2. \quad \square\end{aligned}$$

**Proof of Corollary 4.2.** We have  $\sqrt{n}(\hat{\mu} - \mu_0) = \sum_{S \in \mathcal{S}} [\hat{\omega}_S\sqrt{n}(\hat{\mu}_S - \mu_0)]$  because the weights sum to one. By Corollary 3.1,

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

and by the assumptions of this Corollary we find that  $\hat{\omega}_S \rightarrow_d \varphi_S(\tau, M)$  for each  $S \in \mathcal{S}$ , where  $\varphi_S(\tau, M)$  is a function of  $M$  and constants only. Hence  $\hat{\omega}_S$  and  $\sqrt{n}(\hat{\mu}_S - \mu_0)$  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 4.3.** By Theorem 3.2  $\sqrt{n}(\hat{\beta}(\omega) - \beta) \xrightarrow{d} N(B(\omega), V(\omega))$  since the weights sum to one, where

$$B(\omega) = \omega\left(\frac{\tau}{\sigma_x^2}\right), \quad V(\omega) = \frac{\sigma_\epsilon^2}{\sigma_x^2}\left[(2\omega^2 - \omega)\left(\frac{\sigma_x^2}{\gamma^2} - 1\right) + \frac{\sigma_x^2}{\gamma^2}\right]$$



and accordingly

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

The preceding expression is a globally convex function of  $\omega$ . Taking the first order condition and rearranging gives the desired result.  $\square$

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

$$\sqrt{n}[\mathcal{E}_S f_n(\hat{\theta}_S)] = \sqrt{n}[\mathcal{E}_S f_n(\theta_0)] + F_S \sqrt{n}(\hat{\theta}_S - \theta_0) + o_p(1).$$

Since  $\sqrt{n}(\hat{\theta}_S - \theta_0) \rightarrow_p - (F'_S W_S F_S)^{-1} F'_S W_S \sqrt{n}[\mathcal{E}_S f_n(\theta_0)]$ , we have

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

Thus, for estimation using the efficient weighting matrix

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

by Assumption 2.2(h), where  $\hat{\Omega}_S^{-1/2}$  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>33</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  $\Delta_n(S_1, S_2) = \text{MSC}(S_1) - \text{MSC}(S_2)$ . By Theorem 4.1,  $J_n(S) = O_p(1)$ ,  $S \in \mathcal{S}$ , thus

$$\Delta_n(S_1, S_2) = [J_n(S_1) - J_n(S_2)] - [h(p + |S_1|) - h(p + |S_2|)] \kappa_n = O_p(1) - C \kappa_n$$

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:

$$|S_1| > |S_2| \implies \Delta_n(S_1, S_2) \rightarrow -\infty$$

$$|S_1| = |S_2| \implies \Delta_n(S_1, S_2) = O_p(1)$$

$$|S_1| < |S_2| \implies \Delta_n(S_1, S_2) \rightarrow \infty.$$

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

**Proof of Theorem 4.4.** By Theorem 3.1 and Corollary 4.2,

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

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

$$a_{\min}(\Psi M + \tau) = \min\{a(\tau^*): \tau^* \in \mathcal{T}(\Psi M + \tau, \delta)\}$$

$$b_{\max}(\Psi M + \tau) = \max\{b(\tau^*): \tau^* \in \mathcal{T}(\Psi M + \tau, \delta)\}$$

$$\mathcal{T}(\Psi M + \tau, \delta) = \{\tau^*: \Delta(\tau, \tau^*) \leq \chi_q^2(\delta)\}$$

$$\Delta(\tau, \tau^*) = (\Psi M + \tau - \tau^*)'(\Psi \Omega \Psi')^{-1}(\Psi M + \tau - \tau^*).$$

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.  $\square$

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/j.jeconom.2016.07.006>.

## References

- Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: An empirical investigation. *Amer. Econ. Rev.* 91 (5), 1369–1401.
- Andrews, D.W.K., 1988. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* 4 (3), 458–467.
- Andrews, D.W.K., 1992. Generic uniform convergence. *Econometric Theory* 8 (2), 241–257.
- Andrews, D.W.K., 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. *J. Econometrics* 101, 123–164.
- Berger, R.L., Boos, D.D., 1994. P values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* 89 (427), 1012–1016.
- Berkowitz, D., Caner, M., Fang, Y., 2008. Are nearly exogenous instruments reliable? *Econom. Lett.* 108, 20–23.
- Berkowitz, D., Caner, M., Fang, Y., 2012. The validity of instruments revisited. *J. Econometrics* 166, 255–266.
- Caner, M., 2009. Lasso-type GMM estimator. *Econometric Theory* 25, 270–290.
- Caner, M., 2014. Near exogeneity and weak identification in generalized empirical likelihood estimators: Many moment asymptotics. *J. Econometrics* 182, 247–268.
- Caner, M., Han, X., Lee, Y., 2015. Adaptive elastic net GMM estimation with many invalid moment conditions: Simultaneous model and moment selection. *J. Bus. Econom. Statist.* in press.
- Carstensen, K., Gundlach, E., 2006. The primacy of institutions reconsidered: Direct income effects of malaria prevalence. *World Bank Econ. Rev.* 20 (3), 309–339.
- Chen, X., Jacho-Chavez, 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>.
- Cheng, X., Liao, Z., Shi, R., October 2014. Uniform asymptotic risk of averaging gmm estimator robust to misspecification, Working Paper.
- 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. *J. Amer. Statist. Assoc.* 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. In: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge.

<sup>33</sup> See Hall (2005), Chapter 3.

- Demetrescu, M., Hassler, U., Kuzin, V., 2011. Pitfalls of post-model-selection testing: Experimental quantification. *Empir. Econom.* 40, 359–372.
- Donald, S.G., Imbens, G.W., Newey, W.K., 2009. Choosing instrumental variables in conditional moment restriction models. *J. Econometrics* 152, 28–36.
- Donald, S.G., Newey, W.K., 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. *J. Monet. Econ.* 50, 3–39.
- 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. *J. Appl. Econometrics* 27, 1138–1160.
- Hall, A.R., 2005. Generalized Method of Moments, in: *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 Rev.* 22, 269–288.
- Hansen, B.E., 2015. A Stein-like 2SLS estimator. *Econometric Reviews* (forthcoming).
- Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. *J. Amer. Statist. Assoc.* 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.
- Judge, G.G., Mittelhammer, R.C., 2007. Estimation and inference in the case of competing sets of estimating equations. *J. 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. *J. Amer. Statist. Assoc.* 101 (474), 819–829.
- Kinal, T.W., 1980. The existence of moments of k-class estimators. *Econometrica* 48 (1), 241–249.
- Kuersteiner, G., Okui, R., 2010. Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78 (2), 679–718.
- 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. *J. Econometrics* 142, 201–211.
- Leeb, H., Pötscher, B.M., 2009. Model selection. In: *Handbook of Financial Time Series*. Springer.
- Leeb, H., Pötscher, B.M., 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., 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. *J. Amer. Statist. Assoc.* 80 (390), 362–368.
- McCloskey, A., 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. *J. Econometrics* 29, 229–256.
- Newey, W.K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing, vol. IV. Elsevier Science, pp. 2111–2245 (Chapter 36).
- 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.
- Rodrik, D., Subramanian, A., Trebbi, F., 2004. Institutions rule: The primacy of institutions over geography and integration in economic development. *J. Econ. Growth* 9, 131–165.
- Sachs, J.D., 2003. Institutions don't rule: Direct effects of geography on per capita income, NBER Working Paper No. 9490. <http://www.nber.org/papers/w9490>.
- Schorfheide, F., 2005. VAR forecasting under misspecification. *J. Econometrics* 128, 99–136.
- Silvapulle, M.J., 1996. A test in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* 91 (436), 1690–1693.
- Xiao, Z., 2010. The weighted method of moments approach for moment condition models. *Econom. Lett.* 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.