Job Market Paper

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

Francis J. DiTraglia[a]

[a]*Faculty of Economics, University of Cambridge*

Abstract

In finite samples, the use of a slightly invalid but highly relevant instrument can substantially reduce mean-squared error (MSE). Building on this observation, I propose a moment selection criterion for GMM in which over-identifying restrictions are chosen based on the MSE of their associated estimators rather than their validity: the focused moment selection criterion (FMSC). I then show how the asymptotic framework used to derive the FMSC can be employed to address the problem of inference post-moment selection. Treating post-selection estimators as a special case of moment-averaging, in which estimators based on different moment sets are given data-dependent weights, I propose a simulation-based procedure to construct valid confidence intervals. In a Monte Carlo experiment for 2SLS estimation, the FMSC performs well relative to alternatives suggested in the literature, and the simulation-based procedure achieves its stated minimum coverage. I conclude with an empirical example examining the effect of instrument selection on the estimated relationship between malaria transmission and economic development.

Keywords: Moment selection, GMM estimation, Model averaging, Focused Information Criterion, Post-selection estimators
JEL: C21, C26, C52

1. Introduction

For consistent estimates, instrumental variables must be valid and relevant: correlated with the endogenous regressors but uncorrelated with the error term. In finite samples, however, the use of an invalid but sufficiently relevant instrument can improve inference, reducing estimator variance by far more than bias is increased. Building on this observation, I propose a new 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

☆I thank Gerda Claeskens, Toru Kitagawa, Hashem Pesaran, Richard J. Smith, Stephen Thiele, Melvyn Weeks, as well as seminar participants at Cambridge, Oxford, and the 2011 Econometric Society European Meetings for their many helpful comments and suggestions. I thank Kai Carstensen for providing data for my empirical example.

Email address: fjd26@cam.ac.uk (Francis J. DiTraglia)
URL: http://www.ditraglia.com (Francis J. DiTraglia)

Current Version: January 20, 2012

to yield the smallest mean squared error (MSE) GMM estimator of a user-specified target parameter. I derive FMSC using asymptotic mean squared error (AMSE) to approximate finite-sample MSE. To ensure that AMSE remains finite, I employ a drifting asymptotic framework in which mis-specification, while present for any fixed sample size, vanishes in the limit. 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 FMSC is to trade off these two effects in the limit as an approximation to finite sample behavior. While estimating asymptotic variance is straightforward, even under local mis-specification, estimating asymptotic bias requires over-identifying information. I consider a setting in which two blocks of moment conditions are available: one that is assumed correctly specified, and another that may not be. When the correctly specified block identifies the model, I derive an asymptotically unbiased estimator of AMSE: the FMSC. When this is not the case, it remains possible to use the AMSE framework to carry out a sensitivity analysis.

Still employing the local mis-specification assumption, I show how the ideas used to derive FMSC can be applied to the important problem of inference post-moment selection. Because they use the same data twice, first to choose a moment set and then to carry out estimation, post-selection estimators are randomly weighted averages of many individual estimators. While this is typically ignored in practice, its effects can be dramatic: coverage probabilities of traditional confidence intervals are generally far too low, even for consistent moment selection. I treat post-selection estimators as a special case of moment averaging: combining estimators based on different moment sets with data-dependent weights. By deriving the limiting distribution of moment average estimators, I propose a simulation-based procedure for constructing valid confidence intervals. This technique can be used to correct confidence intervals for a number of moment selection procedures including FMSC.

While the methods described here apply to any model estimated by GMM, subject to standard regularity conditions, I focus on their application to linear instrumental variables (IV) models. In simulations for two-stage least squares (2SLS), FMSC performs well relative to alternatives suggested in the literature. Further, the procedure for constructing valid confidence intervals achieves its stated minimum coverage, even in situations where instrument selection leads to highly non-normal sampling distributions. I conclude with an empirical application from development economics, exploring the effect of instrument selection on the estimated relationship between malaria transmission and income.

My approach to moment selection under mis-specification is inspired by the focused information criterion of Claeskens and Hjort (2003), a model selection criterion for models estimated by maximum likelihood. Like them, I allow for mis-specification and use AMSE to approximate small-sample MSE in a drifting asymptotic framework. In contradistinction, however, I consider moment rather than model selection, and general GMM estimation rather than maximum likelihood.

The existing literature on moment selection under mis-specification is comparatively small. Andrews (1999) proposes a family of moment selection criteria for GMM by adding a penalty term to the J-test statistic. Under an identification assumption and certain restrictions on the form of the penalty, these criteria consistently select all correctly specified moment conditions in the limit. Andrews and Lu (2001) extend this work to allow simulta-

neous GMM moment and model selection, while Hong et al. (2003) derive analogous results for generalized empirical likelihood. More recently, Liao (2010) proposes a shrinkage procedure for simultaneous GMM moment selection and estimation. Given a set of correctly specified moment conditions that identifies the model, this method consistently chooses all valid conditions from a second set of potentially mis-specified conditions. In contrast to these proposals, which examine only the validity of the moment conditions under consideration, the FMSC balances validity against relevance to minimize MSE. The only other proposal from the literature to consider both validity and relevance in moment selection is a suggestion by Hall and Peixe (2003) to combine their canonical correlations information criterion (CCIC) – a relevance criterion that seeks to avoid including redundant instruments – with Andrews' GMM moment selection criteria. This procedure, however, merely seeks to avoid including redundant instruments after eliminating invalid ones: it does not allow for the intentional inclusion of a slightly invalid but highly relevant instrument to reduce MSE. The idea of choosing instruments to minimize MSE is shared by the procedures in Donald and Newey (2001) and Donald et al. (2009). Kuersteiner and Okui (2010) also aim to minimize MSE but, rather than choosing a particular instrument set, suggest averaging over the first-stage predictions implied by many instrument sets and using this average in the second stage. Unlike FMSC, 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.

The literature on post-selection, or "pre-test" estimators is vast. Leeb and Pötscher (2005, 2009) give a theoretical overview, while Demetrescu et al. (2011) illustrate the practical consequences via a simulation experiment. There are several proposals to construct valid confidence intervals post-model selection, including Kabaila (1998), Hjort and Claeskens (2003) and Kabaila and Leeb (2006). To my knowledge, however, this is the first paper to examine the problem specifically from the perspective of moment selection. The approach adopted here, treating post-moment selection estimators as a specific example of moment averaging, is adapted from the frequentist model average estimators of Hjort and Claeskens (2003). Another paper that considers weighting GMM estimators based on different moment sets is Xiao (2010). While Xiao combines estimators based on valid moment conditions to achieve a minimum variance estimator, I combine estimators based on potentially invalid conditions to minimize MSE.

The remainder of the paper is organized as follows. Section 2 describes the local misspecification framework and gives the main limiting results used later in the paper. Section 3 derives FMSC as an asymptotically unbiased estimator of AMSE, presents specialized results for 2SLS, and examines their performance in a Monte Carlo experiment. Section 4 describes a simulation-based procedure to construct valid confidence intervals for moment average estimators and examines its performance in a Monte Carlo experiment. Section 5 presents the empirical application and Section 6 concludes. Proofs, along with supplementary figures and tables, appear in the Appendix.

2. Notation and Asymptotic Framework

Let $f(\cdot,\cdot)$ be a $(p+q)$-vector of moment functions of a random vector Z and r-dimensional parameter vector θ, partitioned according to $f(\cdot,\cdot) = (g(\cdot,\cdot)', h(\cdot,\cdot)')'$ where $g(\cdot,\cdot)$ and $h(\cdot,\cdot)$

are p- and q-vectors of moment functions. The moment condition associated with $g(\cdot,\cdot)$ is assumed to be correct whereas that associated with $h(\cdot,\cdot)$ is locally mis-specified. To be more precise,

Assumption 2.1 (Local Mis-Specification)**.**

$$\mathbb{E}[f(Z_{ni},\theta)] = \mathbb{E}\left[\begin{array}{c} g(Z_{ni},\theta_0) \\ h(Z_{ni},\theta_0) \end{array}\right] = \left[\begin{array}{c} 0 \\ \tau/\sqrt{n} \end{array}\right]$$

where τ is an unknown, q-dimensional vector of constants.

For any fixed sample size n, the expectation of h evaluated at the true parameter value value θ_0 depends on the unknown constant vector τ. Unless all components of τ 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. Local mis-specification is used here as a device to ensure that squared asymptotic bias is of the same order as asymptotic variance.

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

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(Z_{ni},\theta) = \left[\begin{array}{c} g_n(\theta) \\ h_n(\theta) \end{array}\right] = \left[\begin{array}{c} n^{-1}\sum_{i=1}^{n} g(Z_{ni},\theta) \\ n^{-1}\sum_{i=1}^{n} h(Z_{ni},\theta) \end{array}\right] \tag{2.1}$$

In particular, g_n is the sample analogue of the correctly specified moment conditions and h_n that of the mis-specified moment conditions. To describe the two estimators that will play an important role in later results, let $\widetilde{W}$ be a $(q+p)\times(q+p)$, positive semi-definite weighting matrix

$$\widetilde{W} = \left[\begin{array}{cc} \widetilde{W}_{gg} & \widetilde{W}_{gh} \\ \widetilde{W}_{hg} & \widetilde{W}_{hh} \end{array}\right] \tag{2.2}$$

partitioned conformably to the partition of $f(Z,\theta)$ by $g(Z,\theta)$ and $h(Z,\theta)$.

The *valid* estimator uses only those moment conditions known to be correctly specified:

$$\widehat{\theta}_v = \underset{\theta\in\Theta}{\arg\min}\ g_n(\theta)'\widetilde{W}_{gg}\, g_n(\theta) \tag{2.3}$$

For estimation based on g alone to be possible, we must have $p \geq r$. With the exception of Section 3.2, this assumption is maintained throughout.

The *full* estimator uses all moment functions, including the possibly invalid ones contained in h

$$\widehat{\theta}_f = \underset{\theta\in\Theta}{\arg\min}\ f_n(\theta)'\widetilde{W}\, f_n(\theta) \tag{2.4}$$

For this estimator to be feasible, we must have $(p+q) \geq r$. Note that the same weights are used for g in both the valid and full estimation criteria. Although not strictly necessary, this simplifies notation and is appropriate for the efficient GMM estimator.

To consider the limit distributions of $\widehat{\theta}_f$ and $\widehat{\theta}_v$ we require some further notation. For simplicity, assume that the triangular array $\{Z_{ni}\}_{i=1}^{n}$ is asymptotically stationary and denote by Z its almost-sure limit, i.e. $\lim_{n\to\infty}\mathbb{P}\{Z_{ni} = Z\} = 1$ for all i. By Assumption 2.1,

$\mathbb{E}[f(Z,\theta_0)] = 0$. Define

$$F = \begin{bmatrix} G \\ H \end{bmatrix} = \mathbb{E}\begin{bmatrix} \nabla_\theta\, g(Z,\theta_0) \\ \nabla_\theta\, h(Z,\theta_0) \end{bmatrix} \tag{2.5}$$

and

$$\Omega = Var\begin{bmatrix} g(Z,\theta_0) \\ h(Z,\theta_0) \end{bmatrix} = \begin{bmatrix} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{bmatrix} \tag{2.6}$$

Notice that each of these expressions involves the limit random variable Z rather than Z_{ni}, i.e. the corresponding expectations are taken with respect to a distribution for which all moment conditions are correctly specified.

The following high level assumptions are sufficient for the consistency and asymptotic normality of the full and valid estimators.

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

(a) θ_0 *lies in the interior of* Θ*, a compact set*

(b) $\widetilde{W} \to_p W$*, a positive definite matrix*

(c) $W\mathbb{E}[f(Z,\theta)] = 0$ *and* $W_{gg}\mathbb{E}[g(Z,\theta)] = 0$ *if and only if* $\theta = \theta_0$

(d) $\mathbb{E}[f(Z,\theta)]$ *is continuous on* Θ

(e) $\sup_{\theta\in\Theta} \|f_n(\theta) - \mathbb{E}[f(Z,\theta)]\| \to_p 0$

(f) f *is almost surely differentiable in an open neighborhood* $\mathcal{B}$ *of* θ_0

(g) $\sup_{\theta\in\mathcal{B}} \|\nabla_\theta f_n(\theta) - F(\theta)\| \to_p 0$

(h) $\sqrt{n} f_n(\theta_0) \to_d \mathcal{N}_{p+q}\left(\begin{bmatrix} 0 \\ \tau \end{bmatrix}, \Omega\right)$

(i) $F'WF$ *and* $G'W_{gg}G$ *are invertible*

Although Assumption 2.2 closely approximates the standard regularity conditions for GMM estimation, establishing primitive conditions for Assumptions 2.2 (d), (e), (g) and (h) is somewhat more involved under local mis-specification. Appendix B provides details for the case where $\{Z_{ni}\}_{i=1}^n$ is iid over i for fixed n. Notice that identification, (c), and continuity, (d), are conditions on the distribution of Z, the limiting random vector to which $\{Z_{ni}\}_{i=1}^n$ converges.

Under Assumptions 2.1 and 2.2 both the valid and full estimators are consistent and asymptotically normal. The full estimator, however, shows an asymptotic bias. Let

$$M = \begin{bmatrix} M_g \\ M_h \end{bmatrix} \sim \mathcal{N}_{p+q}\left(\begin{bmatrix} 0 \\ \tau \end{bmatrix}, \Omega\right) \tag{2.7}$$

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

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

$$\sqrt{n}\left(\widehat{\theta}_v - \theta_0\right) \to_d -[G'W_{gg}G]^{-1}G'W_{gg}M_h$$

and

$$\sqrt{n}\left(\widehat{\theta}_f - \theta_0\right) \to_d -[F'WF]^{-1}F'WM$$

To study moment selection generally, we need to describe the limit behavior of estimators based on any subset of the the moment conditions contained in h. Fortunately, this only requires some additional notation. Define an arbitrary moment set S by the components of h that it includes. We will always include the moment conditions contained in g. Since h is q-dimensional, $S \subseteq \{1, 2, \ldots, q\}$. For $S = \emptyset$, we have the valid moment set; for $S = \{1, 2, \ldots, q\}$, the full moment set. Denote the number of components from h included in S by $|S|$. Let Ξ_S be the $(p + |S|) \times (p + q)$ selection matrix that extracts those elements of a $(p+q)$-vector corresponding to the moment set S: all p of the first $1, \ldots, p$ components and the specified subset of the $p + 1, \ldots, p + q$ remaining components. Accordingly, define the GMM estimator based on moment set S by

$$\widehat{\theta}_S = \arg\min_{\theta\in\Theta} \ [\Xi_S f_n(\theta)]' \left[\Xi_S \widetilde{W} \Xi_S'\right] \ [\Xi_S f_n(\theta)] \tag{2.8}$$

To simplify the notation let $F_S = \Xi_S F$, $W_S = \Xi_S W \Xi_S'$, $M_S = \Xi_S M$ and $\Omega_S = \Xi_S \Omega \Xi_S'$. Define

$$K_S = [F_S' W_S F_S]^{-1} F_S' W_S. \tag{2.9}$$

Then, by an argument nearly identical to the proof of Theorem 2.2, we have the following.

Corollary 2.1 (Estimators for Arbitrary Moment Sets)**.** *Assume that*

(a) $W_S \Xi_S \mathbb{E}[f(Z, \theta)] = 0$ *if and only if* $\theta = \theta_0$*, and*

(b) $F_S' W_S F_S$ *is invertible.*

Then, under Assumptions 2.1 and 2.2, $\sqrt{n}(\widehat{\theta}_S - \theta_0) \to_d -K_S M_S$*.*

Conditions (a) and (b) from Corollary 2.1 are analogous to Assumption 2.2 (c) and (i).

3. The Focused Moment Selection Criterion

3.1. The General Case

FMSC chooses among the potentially invalid moment conditions contained in h to minimize estimator AMSE for a target parameter. Denote this target parameter by μ, a real-valued, almost-surely continuous function of the parameter vector θ. Further, define the GMM estimator of μ based on $\widehat{\theta}_S$ by $\widehat{\mu}_S = \mu(\widehat{\theta}_S)$ and the true value of μ by $\mu_0 = \mu(\theta_0)$. Applying the delta method to Corollary 2.1 gives the AMSE of $\widehat{\mu}_S$.

Corollary 3.1 (AMSE of Target Parameter)**.** *Under the hypotheses of Corollary 2.1,*

$$\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right) \to_d -\nabla_\theta \mu(\theta_0)' K_S M_S.$$

In particular

$$AMSE(\widehat{\mu}_S) = \nabla_\theta \mu(\theta_0)' K_S \Xi_S \left\{ \left[\begin{array}{cc} 0 & 0 \\ 0 & \tau\tau' \end{array} \right] + \Omega \right\} \Xi_S' K_S' \nabla_\theta \mu(\theta_0).$$

For the full and valid moment sets, we have

$$\begin{aligned} K_f &= [F'WF]^{-1} F'W && (3.1) \\ K_v &= [G'W_{gg}G]^{-1} G'W_{gg} && (3.2) \\ \Xi_f &= \mathbf{I}_{p+q} && (3.3) \\ \Xi_v &= \left[\begin{array}{cc} \mathbf{I}_p & \mathbf{0}_{p\times q} \end{array} \right] && (3.4) \end{aligned}$$

Thus, the valid estimator $\widehat{\mu}_v$ of μ has zero asymptotic bias while the full estimator, $\widehat{\mu}_f$, inherits a bias from each component of τ. Typically, however, $\widehat{\mu}_f$ has the smallest asymptotic variance. In particular, the usual proof that adding moment conditions cannot increase asymptotic variance under efficient GMM continues to hold under local mis-specification, because all moment conditions are correctly specified in the limit.[1] Estimators based on other moment sets lie between these two extremes: the precise nature of the bias-variance tradeoff is governed by the size of the respective components of τ and the projections implied by the matrices K_S. Thus, local mis-specification gives an asymptotic analogue of the finite-sample observation that adding a slightly invalid but highly relevant instrument can decrease MSE.

To use this framework for moment selection, we need to construct estimators of the unknown quantities: θ_0, K_S, Ω, and τ. Under local mis-specification, the estimator of θ under *any* moment set is consistent. In particular, Theorem 2.1 establishes that both the valid and full estimators yield a consistent estimate of θ_0. Recall that $K_S = [F_S' W_S F_S]^{-1} F_S' W_S \Xi_S$. Now, Ξ_S is known because it is simply the selection matrix defining moment set S. 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 Ω are readily available under local mis-specification, although the precise form depends on the situation. Section 3.3 considers this point in more detail for 2SLS and the case of micro-data.

The only remaining unknown is τ. Estimating this quantity, however, is more challenging. The local mis-specification framework is essential for making meaningful comparisons of AMSE, but prevents us from consistently estimating the asymptotic bias parameter. When the correctly specified moment conditions in g identify θ_0, however, we can construct an asymptotically unbiased estimator $\widehat{\tau}$ of τ by substituting $\widehat{\theta}_v$, the estimator of θ_0 that uses only correctly specified moment conditions, into h_n, the sample analogue of the mis-specified moment conditions. That is,

$$\widehat{\tau} = \sqrt{n} h_n(\widehat{\theta}_v). \tag{3.5}$$

[1] See, for example Hall (2005, chapter 6).

Returning to Corollary 3.1, we see that it is $\tau\tau'$ rather than τ that enters the expression for AMSE. Although $\widehat{\tau}$ is an asymptotically unbiased estimator of τ, the limiting expectation of $\widehat{\tau}\widehat{\tau}'$ is not $\tau\tau'$ but rather $\tau\tau' + \Psi\Omega\Psi'$. To obtain an asymptotically unbiased estimator of $\tau\tau'$ we must subtract a consistent estimator of $\Psi\Omega\Psi'$ from $\widehat{\tau}\widehat{\tau}'$.

Theorem 3.1 (Asymptotic Distribution of $\widehat{\tau}$)**.** *Suppose that $p \geq r$. Then,*

$$\widehat{\tau} = \sqrt{n} h_n(\widehat{\theta}_v) \to_d \Psi M$$

where $\Psi = \left[\begin{array}{cc} -HK_v & \mathbf{I}_q \end{array} \right]$. *Therefore* $\Psi M \sim \mathcal{N}_q(\tau, \Psi\Omega\Psi')$.

Corollary 3.2 (Asymptotically Unbiased Estimator of $\tau\tau'$)**.** *Let $\widehat{\Omega}$ and $\widehat{\Psi}$ be consistent estimators of Ω and Ψ. Then,*

$$\widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}' \to_d \Psi \left(MM' - \Omega \right) \Psi'$$

i.e. $\widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}'$ provides an asymptotically unbiased estimator of $\tau\tau'$.

Therefore,

$$\text{FMSC}_n(S) = \nabla_\theta \mu(\widehat{\theta})' \widehat{K}_S \Xi_S \left\{ \left[\begin{array}{cc} 0 & 0 \\ 0 & \widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}' \end{array} \right] + \widehat{\Omega} \right\} \Xi_S' \widehat{K}_S' \nabla_\theta \mu(\widehat{\theta}) \tag{3.6}$$

provides an asymptotically unbiased estimator of AMSE.

3.2. Digression: The Case of $r > p$

When $r > p$, the dimension of the parameter vector θ exceeds that of the moment function vector. Thus θ_0 is not estimable by $\widehat{\theta}_v$ so $\widehat{\tau}$ is not a feasible estimator of τ. A naïve approach to this problem would be to substitute another consistent estimator of θ_0, e.g. $\widehat{\theta}_f$, and proceed analogously. Unfortunately, this approach fails. To understand why, consider the case in which all moment conditions are potentially invalid so the full moment set is h. By a slight variation in the argument used in the proof of Theorem 3.1 we have $\sqrt{n} h_n(\widehat{\theta}_f) \to_d \Gamma \times \mathcal{N}_q(\tau, \Omega)$, where

$$\Gamma = \mathbf{I}_q - H \left(H'WH \right)^{-1} H'W \tag{3.7}$$

The mean, $\Gamma\tau$, of the resulting limit distribution does not in general equal τ. Because Γ has rank $q - r$ we cannot pre-multiply by its inverse to extract an estimate of τ. Intuitively, $q - r$ over-identifying restrictions are insufficient to estimate the q-vector τ.

Thus, τ is not identified unless we have a minimum of r valid moment conditions. However, the limiting distribution of $\sqrt{n} h_n(\widehat{\theta}_f)$ partially identifies τ even when we have no valid moment conditions at our disposal. A combination of this information with prior restrictions on the magnitude of the components of τ allows the use of the FMSC framework to carry out a sensitivity analysis when $r > p$. For example, the worst-case estimate of AMSE over values of τ in the identified region could still allow certain moment sets to be ruled out. This idea shares certain similarities with Kraay (2010) and Conley et al. (2010), two recent papers that suggest methods for evaluating the robustness of conclusions drawn from IV regressions when the instruments used may be invalid.

3.3. The FMSC for 2SLS Instrument Selection

This section specializes FMSC to a case of particular applied interest: instrument selection for 2SLS in a micro-data setting. The expressions given here are used in the simulation studies and empirical example that appear later in the paper.

Consider a linear IV regression model with response variable y_{ni}, regressors $\mathbf{x}_{ni}$, valid instruments $\mathbf{z}_{ni}^{(1)}$ and potentially invalid instruments $\mathbf{z}_{ni}^{(2)}$. Define $\mathbf{z}_{ni} = (\mathbf{z}_{ni}^{(1)}, \mathbf{z}_{ni}^{(2)})'$. We assume that $\{(y_{ni}, \mathbf{x}_{ni}', \mathbf{z}_{ni}')'\}_{i=1}^{n}$ is iid across i for fixed sample size n, but allow the distribution to change with n. In this case, the local mis-specification framework given in Assumption 2.1 becomes

$$\mathbb{E}\left[\begin{array}{c} \mathbf{z}_{ni}^{(1)}\left(y_{ni} - \mathbf{x}_{ni}'\theta_0\right) \\ \mathbf{z}_{ni}^{(2)}\left(y_{ni} - \mathbf{x}_{ni}'\theta_0\right) \end{array}\right] = \left[\begin{array}{c} \mathbf{0} \\ \tau/\sqrt{n} \end{array}\right]. \tag{3.8}$$

Stacking observations, let $X = (\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nn})'$, $y = (y_{n1}, \ldots, y_{nn})'$, $Z_1 = (\mathbf{z}_{n1}^{(1)}, \ldots, \mathbf{z}_{nn}^{(1)})'$, $Z_2 = (\mathbf{z}_{n1}^{(2)}, \ldots, \mathbf{z}_{nn}^{(2)})'$, and $Z = (Z_1, Z_2)$. Further define $u_{ni}(\theta) = y_{ni} - \mathbf{x}_{ni}'\theta$ and $u(\theta) = y - X\theta$. The 2SLS estimator of θ_0 under instrument set S is given by

$$\widehat{\theta}_S = \left[X'Z_S\left(Z_S'Z_S\right)^{-1}Z_S'X\right]^{-1}X'Z_S\left(Z_S'Z_S\right)^{-1}Z_S'y \tag{3.9}$$

where $Z_S = Z\Xi_S'$. Similarly, the full and valid estimators are

$$\widehat{\theta}_f = \left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}X'Z\left(Z'Z\right)^{-1}Z'y \tag{3.10}$$

$$\widehat{\theta}_v = \left[X'Z_1\left(Z_1'Z_1\right)^{-1}Z_1'X\right]^{-1}X'Z_1\left(Z_1'Z_1\right)^{-1}Z_1'y \tag{3.11}$$

Let $(\mathbf{x}', \mathbf{z}')'$ be the almost-sure limit of $\{(\mathbf{x}_{ni}', \mathbf{z}_{ni}')'\}_{i=1}^{n}$ as $n \to \infty$ and define $\mathbf{z}_S = \Xi_S\mathbf{z}$. Then, the matrix K_S defined in 2.9 becomes

$$K_S = -\left(\mathbb{E}\left[\mathbf{x}\mathbf{z}_S'\right]\left(\mathbb{E}[\mathbf{z}_S\mathbf{z}_S']\right)^{-1}\mathbb{E}\left[\mathbf{z}_S'\mathbf{x}\right]\right)^{-1}\mathbb{E}\left[\mathbf{x}\mathbf{z}_S'\right]\left(\mathbb{E}[\mathbf{z}_S\mathbf{z}_S']\right)^{-1}. \tag{3.12}$$

Because observations are iid for fixed n,

$$\Omega = \lim_{n\to\infty} Var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\mathbf{z}_{ni}u_{ni}(\theta_0)\right) = \lim_{n\to\infty} Var\left[\mathbf{z}_{ni}u_{ni}(\theta_0)\right] \tag{3.13}$$

This expression allows for conditional but not unconditional heteroscedasticity.

To use the FMSC for instrument selection, we first need an estimator of K_S for each moment set under consideration, e.g.

$$\widehat{K}_S = n\left[X'Z_S\left(Z_S'Z_S\right)^{-1}Z_S'X\right]^{-1}X'Z_S\left(Z_S'Z_S\right)^{-1} \tag{3.14}$$

which is consistent for K_S under Assumption 2.2. To estimate Ω for all but the valid

instrument set, I employ the centered, heteroscedasticity-consistent estimator

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' u_i(\widehat{\theta}_f)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i u_i(\widehat{\theta}_f)\right)\left(\frac{1}{n}\sum_{i=1}^{n} u_i(\widehat{\theta}_f)\mathbf{z}_i'\right).$$

Centering allows moment functions to have non-zero means. While the local mis-specification framework implies that these means tend to zero in the limit, they are non-zero for any fixed sample size. Centering accounts for this fact, and thus provides added robustness.

Since the valid estimator $\widehat{\theta}_v$ has no asymptotic bias, the AMSE of any target parameter based on $\widehat{\theta}_v$ equals asymptotic variance. Rather than using the $(p \times p)$ upper left sub-matrix of $\widehat{\Omega}$ to estimate this quantity, I use

$$\widetilde{\Omega}_{11} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{1i} \mathbf{z}_{1i}' u_i(\widehat{\theta}_v)^2. \tag{3.15}$$

This estimator imposes the assumption that all instruments in Z_1 are valid so that no centering is needed, and thus should be more precise. A robust estimator of $\nabla_\theta \mu(\theta_0)$ is provided by $\nabla_\theta \mu(\widehat{\theta}_{Valid})$. The only remaining quantity needed for FMSC is the asymptotically unbiased estimator $\widehat{\tau}\widehat{\tau} - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}$ of $\tau\tau'$ (see Theorem 3.1 and Corollary 3.2). For 2SLS, $\widehat{\tau} = n^{-1/2}\, Z_2'\, u(\hat{\theta}_v)$ while

$$\widehat{\Psi} = \left[\begin{array}{cc} -n^{-1} Z_2' X \widehat{K}_v & \mathbf{I} \end{array}\right]. \tag{3.16}$$

3.4. Simulation Study

In this section I evaluate the performance of FMSC in a simple setting: instrument selection for 2SLS. The simulation setup is as follows. For $i = 1, 2, \ldots, n$

$$\begin{aligned} y_i &= 0.5 x_i + u_i & (3.17)\\ x_i &= 0.1(z_{1i} + z_{2i} + z_{3i}) + \gamma w_i + \epsilon_i & (3.18) \end{aligned}$$

where $(u_i, \epsilon_i, w_i)' \sim$ iid $\mathcal{N}(0, \mathcal{V})$ with

$$\mathcal{V} = \left[\begin{array}{ccc} 1 & 0.5 - \gamma\rho & \rho \\ 0.5 - \gamma\rho & 1 & 0 \\ \rho & 0 & 1 \end{array}\right] \tag{3.19}$$

independently of $(z_{1i}, z_{2i}, z_{3i}) \sim \mathcal{N}(0, \mathbf{I})$. This design keeps the endogeneity of x fixed, $Cov(x, u) = 0.5$, while allowing the validity and relevance of w to vary according to

$$\begin{aligned} Cov(w, u) &= \rho & (3.20)\\ Cov(w, x) &= \gamma & (3.21) \end{aligned}$$

The instruments z_1, z_2, z_3 are valid and relevant: they have first-stage coefficients of 0.1 and are uncorrelated with the second stage error u.

Our goal is to estimate the effect of x on y with minimum MSE by choosing between two estimators: the valid estimator that uses only z_1, z_2, and z_3 as instruments, and the full

Table 1: Difference in RMSE between the estimator including w (full) and the estimator excluding it (valid) over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument. Negative values indicate that including w gives a smaller RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.01	0.00	0.02	0.07	0.13	0.18	0.25	0.31	0.39
0.1	-0.06	0.00	0.09	0.19	0.30	0.42	0.53	0.65	0.79
0.2	-0.10	-0.04	0.07	0.19	0.32	0.46	0.58	0.72	0.86
0.3	-0.14	-0.09	0.01	0.12	0.24	0.36	0.48	0.61	0.72
0.4	-0.17	-0.12	-0.03	0.06	0.16	0.26	0.36	0.46	0.57
0.5	-0.19	-0.15	-0.07	0.01	0.10	0.19	0.27	0.34	0.45
0.6	-0.20	-0.17	-0.10	-0.03	0.04	0.11	0.19	0.26	0.34
0.7	-0.21	-0.18	-0.13	-0.07	-0.01	0.07	0.14	0.20	0.26
0.8	-0.22	-0.20	-0.15	-0.09	-0.04	0.03	0.09	0.15	0.20
0.9	-0.23	-0.21	-0.16	-0.12	-0.07	-0.01	0.04	0.10	0.14
1.0	-0.25	-0.22	-0.19	-0.13	-0.08	-0.04	0.01	0.06	0.11
1.1	-0.24	-0.22	-0.20	-0.16	-0.10	-0.07	-0.02	0.03	0.07
1.2	-0.26	-0.22	-0.19	-0.16	-0.12	-0.07	-0.05	-0.01	0.03
1.3	-0.29	-0.24	-0.20	-0.17	-0.14	-0.09	-0.06	-0.01	0.02

$\gamma = Cov(w, x)$

estimator that 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 the valid estimator and three for the full estimator. Because the moments of the 2SLS estimator only exist up to the order of over-identification (Phillips, 1980), this ensures that the small-sample MSE is well-defined. All simulations are carried out over a grid of values for (γ, ρ) with 10,000 replications at each point. Estimation is by 2SLS without a constant term, using the expressions from Section 3.3.

Table 1 gives the difference in small-sample root mean squared error (RMSE) between the full and valid estimators for a sample size of 500. Negative values indicate parameter values at which the full instrument set has a lower RMSE. We see that even if $Cov(w, u) \neq 0$, so that w is invalid, including it in the instrument set can dramatically lower RMSE provided that $Cov(w, x)$ is high. In other words, using an invalid but sufficiently relevant instrument can improve our estimates. Tables C.22 and C.23 present the same results for sample sizes of 50 and 100, respectively. For smaller sample sizes the full estimator has the lower RMSE over increasingly large regions of the parameter space. Because a sample size of 500 effectively divides the parameter space into two halves, one where the full estimator has the advantage and one where the valid estimator does, I concentrate on this case. Summary results for smaller sample sizes appear in Table 6.

The FMSC chooses moment conditions to minimize an asymptotic approximation to small-sample MSE in the hope that this will provide reasonable performance in practice. The first question is how often the FMSC succeeds in identifying the instrument set that

Table 2: Correct decision rates for the FMSC in percentage points over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. A correct decision is defined as an instance in which the FMSC identifies the estimator that in fact minimizes small sample MSE, as indicated by Table 1. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

		$\rho = Cov(w, u)$								
	$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w, x)$	0.0	79	61	69	85	91	94	94	95	96
	0.1	82	25	62	91	98	99	99	100	100
	0.2	84	82	46	80	96	99	100	100	100
	0.3	85	85	31	60	82	94	98	99	100
	0.4	84	86	77	42	65	82	92	96	98
	0.5	84	87	82	31	49	68	81	90	95
	0.6	84	88	84	75	38	54	68	80	87
	0.7	85	87	86	80	69	44	57	69	79
	0.8	84	87	86	82	74	36	48	60	71
	0.9	85	87	87	84	78	69	41	52	61
	1.0	85	88	87	85	79	74	35	45	53
	1.1	85	88	88	86	82	76	68	39	48
	1.2	85	88	88	87	84	79	72	65	43
	1.3	86	87	88	88	84	80	75	69	39

minimizes small sample MSE. Table 2 gives the frequency of correct decisions in percentage points made by the FMSC for a sample size of 500. A correct decision is defined as an instance in which the FMSC selects the moment set that minimizes finite-sample MSE as indicated by Table 1. We see that the FMSC performs best when there are large differences in MSE between the full and valid estimators: in the top right and bottom left of the parameter space. The criterion performs less well in the borderline cases along the main diagonal.

Ultimately, the goal of the FMSC is to produce estimators with low MSE. Because the FMSC is itself random, however, using it introduces an additional source of variation. Table 3 accounts for this fact by presenting the RMSE that results from using the estimator chosen by the FMSC. Because these values are difficult to interpret on their own, Tables 4 and 5 compare the realized RMSE of the FMSC to those of the valid and full estimators. Negative values indicate that the RMSE of the FMSC is lower. As we see from Table 4, the valid estimator outperforms the FMSC in the upper right region of the parameter space, the region where the valid estimator has a lower RMSE than the full. This is because the FMSC sometimes chooses the wrong instrument set, as indicated by Table 2. Accordingly, the FMSC performs substantially better in the bottom left of the parameter space, the region where the full estimator has a lower RMSE than the valid. Taken on the whole, however, the potential advantage of using the valid estimator is small: at best it yields an RMSE 0.06 smaller than that of the FMSC. Indeed, many of the values in the top right of the parameter space are zero, indicating that the FMSC performs no worse than the valid estimator. In contrast, the potential advantage of using the FMSC is large: it can yield an RMSE 0.16

Table 3: RMSE of the estimator selected by the FMSC over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

		$\rho = Cov(w,u)$								
$N = 500$		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w,x)$	0.0	0.26	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27
	0.1	0.24	0.26	0.28	0.27	0.27	0.27	0.27	0.27	0.27
	0.2	0.22	0.25	0.30	0.31	0.28	0.27	0.28	0.27	0.27
	0.3	0.20	0.23	0.29	0.32	0.31	0.29	0.28	0.27	0.28
	0.4	0.20	0.22	0.27	0.31	0.32	0.31	0.30	0.30	0.28
	0.5	0.20	0.20	0.25	0.29	0.32	0.32	0.32	0.31	0.29
	0.6	0.19	0.19	0.23	0.27	0.30	0.33	0.33	0.32	0.31
	0.7	0.18	0.19	0.22	0.25	0.28	0.31	0.32	0.33	0.32
	0.8	0.18	0.19	0.21	0.24	0.27	0.30	0.31	0.32	0.32
	0.9	0.18	0.19	0.20	0.23	0.26	0.28	0.30	0.32	0.33
	1.0	0.18	0.18	0.19	0.22	0.25	0.27	0.29	0.30	0.32
	1.1	0.17	0.17	0.19	0.21	0.23	0.25	0.28	0.29	0.31
	1.2	0.17	0.17	0.18	0.20	0.22	0.24	0.26	0.28	0.29
	1.3	0.17	0.17	0.17	0.19	0.21	0.23	0.25	0.27	0.28

smaller than the valid model. The situation is similar for the full estimator only in reverse, as shown in Table 5. The full estimator outperforms the FMSC in the bottom left of the parameter space, while the FMSC outperforms the full estimator in the top right. Again, the potential gains from using the FMSC are large compared to those of the full instrument set: a 0.86 reduction in RMSE versus a 0.14 reduction. Average and worst-case RMSE comparisons between the FMSC and the full and valid estimators appear in Table 6.

I now compare the FMSC to a number of alternative procedures from the literature. Andrews introduces the following GMM analogues of Schwarz's Bayesian Information Criterion (BIC), Akaike's Information Criterion (AIC) and the Hannan-Quinn Information Criterion (HQ):

$$\text{GMM-BIC}(S) = J_n(S) - (p + |S| - r) \log n \tag{3.22}$$

$$\text{GMM-HQ}(S) = J_n(S) - 2.01\,(p + |S| - r) \log\log n \tag{3.23}$$

$$\text{GMM-AIC}(S) = J_n(S) - 2\,(p + |S| - r) \tag{3.24}$$

where $J_n(S)$ is the J-test statistic under moment set S. In each case, we choose the moment set S that minimizes the criterion. Under certain assumptions, the HQ and BIC-type criteria are consistent, they select any and all valid moment conditions with probability approaching one in the limit (w.p.a.1). When calculating the J-test statistic under potential mis-specification, Andrews recommends using a centered covariance matrix estimator and basing estimation on the weighting matrix that would be efficient under the assumption of

Table 4: Difference in RMSE between the estimator selected by the FMSC and the valid estimator (which always excludes w) over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.1	-0.04	-0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
0.2	-0.05	-0.02	0.03	0.03	0.00	0.00	0.00	0.00	0.00
0.3	-0.07	-0.04	0.02	0.04	0.04	0.01	0.01	0.00	0.00
0.4	-0.08	-0.05	0.00	0.04	0.05	0.04	0.03	0.02	0.01
0.5	-0.08	-0.07	-0.02	0.02	0.05	0.06	0.05	0.02	0.02
0.6	-0.09	-0.08	-0.04	0.00	0.03	0.04	0.05	0.04	0.04
0.7	-0.09	-0.08	-0.06	-0.03	0.00	0.04	0.05	0.06	0.05
0.8	-0.10	-0.09	-0.07	-0.03	-0.01	0.02	0.04	0.05	0.04
0.9	-0.10	-0.09	-0.08	-0.06	-0.03	0.00	0.02	0.04	0.04
1.0	-0.12	-0.11	-0.10	-0.06	-0.04	-0.02	0.00	0.02	0.04
1.1	-0.11	-0.11	-0.11	-0.09	-0.05	-0.04	-0.02	0.01	0.02
1.2	-0.13	-0.11	-0.11	-0.09	-0.07	-0.04	-0.04	-0.01	0.00
1.3	-0.16	-0.12	-0.11	-0.10	-0.09	-0.05	-0.04	-0.01	0.00

$\gamma = Cov(w, x)$

Table 5: Difference in RMSE between the estimator selected by the FMSC and the full estimator (which always includes w) over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	-0.01	-0.03	-0.07	-0.13	-0.18	-0.25	-0.31	-0.39
0.1	0.02	-0.01	-0.07	-0.18	-0.30	-0.42	-0.53	-0.65	-0.78
0.2	0.05	0.02	-0.04	-0.16	-0.31	-0.46	-0.58	-0.72	-0.86
0.3	0.07	0.05	0.01	-0.08	-0.20	-0.34	-0.47	-0.61	-0.71
0.4	0.09	0.07	0.03	-0.02	-0.11	-0.22	-0.33	-0.44	-0.56
0.5	0.11	0.08	0.05	0.01	-0.05	-0.13	-0.22	-0.32	-0.42
0.6	0.11	0.09	0.07	0.03	-0.01	-0.06	-0.14	-0.22	-0.30
0.7	0.12	0.10	0.07	0.04	0.01	-0.03	-0.08	-0.14	-0.22
0.8	0.13	0.11	0.08	0.05	0.03	0.00	-0.05	-0.10	-0.15
0.9	0.13	0.11	0.08	0.06	0.04	0.01	-0.02	-0.06	-0.10
1.0	0.13	0.11	0.09	0.07	0.05	0.02	-0.01	-0.04	-0.07
1.1	0.13	0.11	0.09	0.07	0.05	0.03	0.01	-0.02	-0.05
1.2	0.14	0.11	0.09	0.07	0.05	0.03	0.02	0.00	-0.03
1.3	0.13	0.12	0.09	0.07	0.05	0.04	0.02	0.00	-0.02

$\gamma = Cov(w, x)$

correct specification. Accordingly, I calculate

$$J_{Full} = \frac{1}{n}\, u(\widehat{\theta}_f)'\, Z\, \widehat{\Omega}^{-1}\, Z'\, u(\widehat{\theta}_f) \tag{3.25}$$

$$J_{Valid} = \frac{1}{n}\, u(\widehat{\theta}_v)'\, Z_1\, \widetilde{\Omega}_{11}^{-1}\, Z_1'\, u(\widehat{\theta}_v) \tag{3.26}$$

for the full and valid instrument sets, respectively, using the formulas from Section 3.3.

Because the Andrews-type criteria only take account of instrument validity, not relevance, Hall and Peixe (2003) suggest combining them with their canonical correlations information criterion (CCIC). The CCIC aims to detect and eliminate redundant instruments, those that add no further information beyond that contained in the other instruments. While including such instruments has no effect on the asymptotic distribution of the estimator, it could lead to poor finite-sample performance. By combining the CCIC with an Andrews-type criterion, the idea is to eliminate invalid instruments and then redundant ones. For the present simulation example, with a single endogenous regressor and no constant term, the CCIC takes the following form (Jana, 2005)

$$\text{CCIC}(S) = n \log\left[1 - R_n^2(S)\right] + h(p + |S|)\mu_n \tag{3.27}$$

where $R_n^2(S)$ is the first-stage R^2 based on instrument set S and $h(p + |S|)\mu_n$ is a penalty term. Specializing these by analogy to the BIC, AIC, and HQ gives

$$\text{CCIC-BIC}(S) = n \log\left[1 - R_n^2(S)\right] + (p + |S| - r) \log n \tag{3.28}$$

$$\text{CCIC-HQ}(S) = n \log\left[1 - R_n^2(S)\right] + 2.01\,(p + |S| - r) \log\log n \tag{3.29}$$

$$\text{CCIC-AIC}(S) = n \log\left[1 - R_n^2(S)\right] + 2\,(p + |S| - r) \tag{3.30}$$

I consider procedures that combine CCIC criteria with the corresponding criterion of Andrews (1999). For the present simulation example, these are as follows:

CC-MSC-BIC: Include w iff doing so minimizes GMM-BIC *and* CCIC-BIC (3.31)

CC-MSC-HQ: Include w iff doing so minimizes GMM-HQ *and* CCIC-HQ (3.32)

CC-MSC-AIC: Include w iff doing so minimizes GMM-AIC *and* CCIC-AIC (3.33)

A less formal but extremely common procedure for moment selection in practice is the downward J-test. In the present context this takes a particularly simple form: if the J-test fails to reject the null hypothesis of correct specification for the full instrument set, use this set for estimation; otherwise, use the valid instrument set. In addition to the moment selection criteria given above, I compare the FMSC to selection by a downward J-test at the 90% and 95% significance levels.

Table 6 compares average and worst-case RMSE over the parameter space given in Table 1 for sample sizes of 50, 100, and 500 observations. For each sample size the FMSC outperforms all other moment selection procedures in both average and worst-case RMSE. The gains are particularly large for smaller sample sizes. Pointwise RMSE comparisons for a sample size of 500 appear in Tables C.24–C.31. These results given here suggest that the FMSC may be of considerable value for instrument selection in practice.

Table 6: Summary of Simulation Results. Average and worst-case RMSE are calculated over the simulation grid from Table 1. All values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications at each point on the grid.

Average RMSE	$N = 50$	$N = 100$	$N = 500$
Valid Estimator	0.69	0.59	0.28
Full Estimator	0.44	0.40	0.34
FMSC	0.47	0.41	0.26
GMM-BIC	0.61	0.52	0.29
GMM-HQ	0.64	0.56	0.29
GMM-AIC	0.67	0.58	0.28
Downward J-test 90%	0.55	0.50	0.28
Downward J-test 95%	0.51	0.47	0.28
CC-MSC-BIC	0.61	0.51	0.28
CC-MSC-HQ	0.64	0.55	0.28
CC-MSC-AIC	0.66	0.57	0.28

Worst-case RMSE	$N = 50$	$N = 100$	$N = 500$
Valid Estimator	0.84	1.06	0.32
Full Estimator	1.04	1.12	1.14
FMSC	0.81	0.74	0.33
GMM-BIC	0.99	0.99	0.47
GMM-HQ	0.97	1.03	0.39
GMM-AIC	0.95	1.04	0.35
Downward J-test 90%	0.99	0.98	0.41
Downward J-test 95%	1.01	1.00	0.46
CC-MSC-BIC	0.86	0.99	0.47
CC-MSC-HQ	0.87	1.03	0.39
CC-MSC-AIC	0.87	1.04	0.35

4. Estimators Post-Selection and Moment Averaging

4.1. The Effects of Moment Selection on Inference

The usual approach to inference post-selection is to state conditions under which traditional distribution theory continues to hold, typically involving an appeal to Lemma 1.1 of Pötscher (1991, p. 168), which states that the limit distributions of an estimator pre- and post-consistent selection are identical. Pötscher (1991, pp. 179–180), however, also states that this result does not hold uniformly over the parameter space. Accordingly, Leeb and Pötscher (2005, p. 22) emphasize that a reliance on the lemma "only creates an illusion of conducting valid inference." In this section, we return to the simulation experiment described in Section 3.4 to briefly illustrate the impact of moment selection on inference.

Figure 4.1 gives the distributions of the valid and full estimators alongside the post-selection distributions of estimators chosen by the GMM-BIC and HQ criteria, defined in Equations 3.22 and 3.23. The distributions are computed by kernel density estimation using $10,000$ replications of the simulation described in Equations 3.17 and 3.18, each with a sample size of 500 and $\gamma = 0.4$, $\rho = 0.2$. For these parameter values the instrument w is relevant but sufficiently invalid that, based on the results of Table 1, we should exclude it. Because GMM-BIC and HQ are consistent procedures, they will exclude any invalid instruments w.p.a.1. A naïve reading of Pötscher's Lemma 1.1 suggests that consistent instrument selection is innocuous, and thus that the post-selection distributions of GMM-BIC and HQ should be close to that of the valid estimator, indicated by dashed lines. This is emphatically not the case: the post-selection distributions are highly non-normal mixtures of the distributions of the valid and full estimators. While Figure 4.1 pertains to only one point in the parameter space, the problem is more general. Tables 7 and 8 give the empirical coverage probabilities of traditional 95% confidence intervals over the full simulation grid. Over the vast majority of the parameter space, empirical coverage probabilities are far lower than the nominal level 0.95. The lack of uniformity is particularly striking. When w is irrelevant, $\gamma = 0$, or valid $\rho = 0$, empirical coverage probabilities are only slightly below 0.95. Relatively small changes in either ρ or γ, however, lead to a large deterioration in coverage.

Because FMSC, GMM-AIC and selection based on a downward J-test at a fixed significance level are not consistent procedures, Lemma 1.1 of Pötscher (1991) is inapplicable. Their behavior, however, is similar to that of the GMM-BIC and HQ (see Figures C.2–C.3 and Tables 9 and C.32–C.34). As this example illustrates, ignoring the effects of moment selection can lead to highly misleading inferences.

4.2. Moment Average Estimators

To account for the effects of moment selection on inference, I extend a framework developed by Hjort and Claeskens (2003) for frequentist model averaging. I treat post-selection estimators as a special case of moment-averaging: combining estimators based on different moment sets using data-dependent weights. Consider an estimator of the form,

$$\widehat{\mu} = \sum_{S \in \mathcal{A}} \widehat{\omega}(S) \widehat{\mu}_S \tag{4.1}$$

where $\widehat{\mu}_S = \mu(\widehat{\theta}_S)$ is an estimator of the target parameter μ under moment set S, $\mathcal{A}$ is the collection of all moment sets under consideration, and the weight function $\widehat{\omega}(\cdot)$ may be

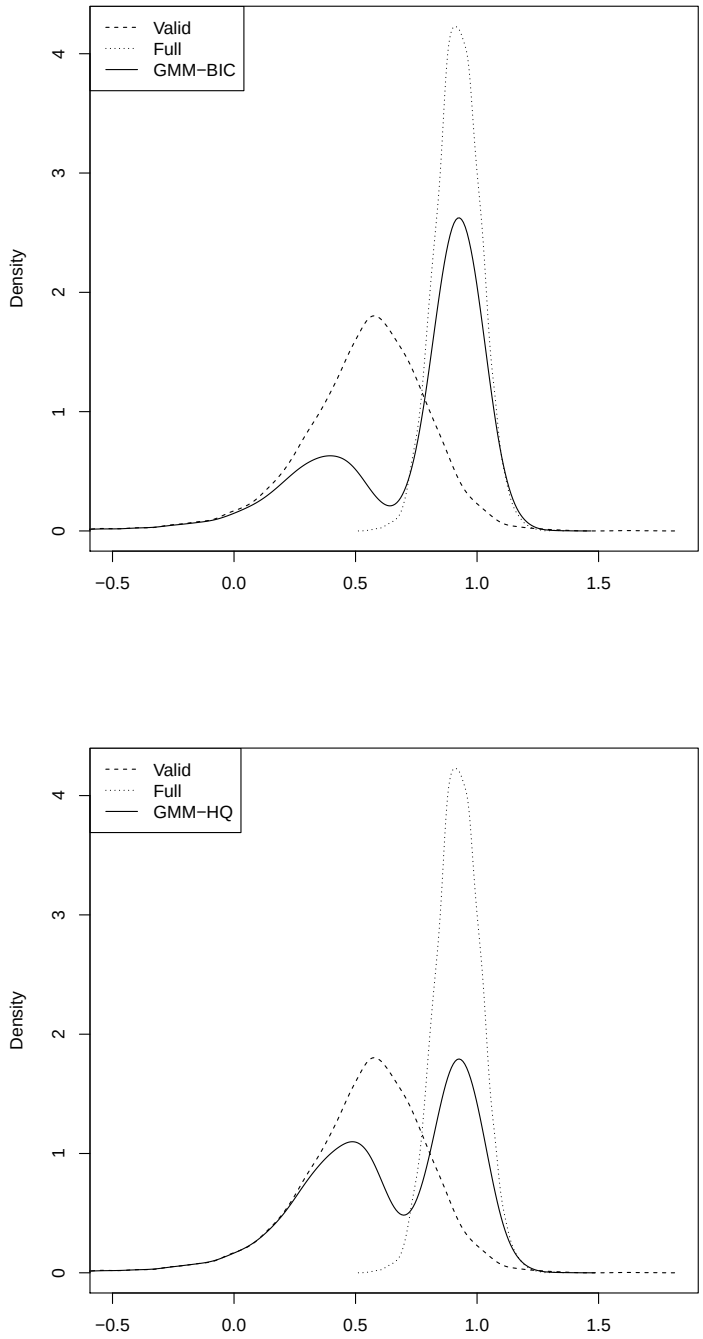

Figure 1: Post-selection distributions for the estimated effect of x on y in Equation 3.17 with $\gamma = 0.4$, $\rho = 0.2$, $N = 500$. The distribution post-GMM-BIC selection appears in the top panel, while the distribution post-GMM-HQ selection appears in the bottom panel. The distribution of the full estimator is given in dotted lines while that of the valid estimator is given in dashed lines in each panel. All distributions are calculated by kernel density estimation based on 10,000 simulation replications generated from Equations 3.17–3.19.

Table 7: Coverage probabilities post-GMM-BIC moment selection of a traditional 95% asymptotic confidence interval for the effect of x on y in Equation 3.17, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

		$\rho = Cov(w, u)$								
$N = 500$		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w, x)$	0.0	0.92	0.92	0.92	0.93	0.92	0.92	0.92	0.92	0.93
	0.1	0.92	0.83	0.77	0.83	0.90	0.92	0.93	0.92	0.92
	0.2	0.93	0.76	0.55	0.57	0.74	0.86	0.89	0.90	0.91
	0.3	0.93	0.75	0.45	0.35	0.50	0.69	0.80	0.85	0.88
	0.4	0.93	0.75	0.40	0.22	0.31	0.48	0.63	0.74	0.80
	0.5	0.93	0.75	0.38	0.18	0.20	0.32	0.46	0.59	0.68
	0.6	0.94	0.76	0.38	0.14	0.14	0.23	0.32	0.43	0.53
	0.7	0.94	0.76	0.37	0.12	0.11	0.16	0.24	0.32	0.42
	0.8	0.93	0.76	0.37	0.11	0.08	0.12	0.18	0.25	0.33
	0.9	0.94	0.75	0.37	0.11	0.07	0.10	0.14	0.19	0.25
	1.0	0.93	0.76	0.37	0.10	0.06	0.08	0.11	0.16	0.20
	1.1	0.93	0.77	0.37	0.10	0.06	0.07	0.10	0.13	0.16
	1.2	0.94	0.77	0.38	0.10	0.05	0.06	0.08	0.11	0.14
	1.3	0.94	0.77	0.38	0.10	0.04	0.05	0.07	0.09	0.12

Table 8: Coverage probabilities post-GMM-HQ moment selection of a traditional 95% asymptotic confidence interval for the effect of x on y in Equation 3.17, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

		$\rho = Cov(w, u)$								
$N = 500$		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w, x)$	0.0	0.92	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	0.1	0.92	0.85	0.84	0.89	0.92	0.93	0.93	0.92	0.93
	0.2	0.92	0.78	0.66	0.74	0.86	0.91	0.91	0.92	0.92
	0.3	0.92	0.76	0.54	0.54	0.69	0.83	0.88	0.90	0.91
	0.4	0.91	0.76	0.47	0.38	0.52	0.69	0.79	0.85	0.88
	0.5	0.91	0.75	0.44	0.30	0.39	0.54	0.67	0.77	0.82
	0.6	0.91	0.76	0.42	0.25	0.29	0.41	0.54	0.64	0.72
	0.7	0.92	0.76	0.40	0.21	0.24	0.33	0.43	0.53	0.63
	0.8	0.91	0.76	0.41	0.19	0.20	0.27	0.36	0.45	0.53
	0.9	0.92	0.75	0.40	0.19	0.17	0.23	0.30	0.38	0.44
	1.0	0.91	0.76	0.40	0.16	0.15	0.20	0.25	0.32	0.38
	1.1	0.91	0.76	0.40	0.16	0.14	0.18	0.23	0.28	0.33
	1.2	0.92	0.76	0.40	0.16	0.13	0.16	0.20	0.25	0.30
	1.3	0.92	0.77	0.41	0.15	0.13	0.15	0.18	0.22	0.27

data-dependent. As above $\mu(\cdot)$ is a $\mathbb{R}$-valued, almost-surely continuous function of θ. When $\widehat{\mu}$ is an indicator function taking on the value one at the moment set S that minimizes some moment selection criterion, $\widehat{\mu}$ is a post-moment selection estimator. More generally, $\widehat{\mu}$ is a moment average estimator.

The limiting behavior of $\widehat{\mu}$ follows almost immediately from Corollary 2.1, which states that asymptotic distribution of $\widehat{\theta}_S$ depends only on K_S and M. Because K_S is a matrix of constants, the random variable M governs the joint limiting behavior of $\widehat{\theta}_S$, $S \in \mathcal{A}$. Under certain conditions on the $\widehat{\omega}(\cdot)$, we can fully characterize the limit distribution of $\widehat{\mu}$.

Assumption 4.1 (Conditions on Weight Functions)**.**

(a) $\sum_{S \in \mathcal{A}} \widehat{\omega}(S) = 1$,

(b) $\widehat{\omega}(\cdot)$ *is almost-surely continuous, and*

(c) $\widehat{\omega}(S) \to_d \omega(M|S)$*, a function of* M *(defined in Theorem 2.2) and constants only.*

Corollary 4.1 (Asymptotic Distribution of Moment-Average Estimators)**.** *Under Assumption 4.1 and the conditions of Corollary 2.1,*

$$\sqrt{n}\left(\widehat{\mu} - \mu_0\right) \to_d \Lambda(\tau) = -\nabla_\theta \mu(\theta_0)' \left[\sum_{S \in \mathcal{A}} \omega(M|S) K_S \Xi_S\right] M.$$

Notice that the limit random variable, denoted $\Lambda(\tau)$, is a randomly weighted average of the multivariate normal vector M. Hence, $\Lambda(\tau)$ is in general non-normal.

Although it restricts the convergence of the weight functions, Assumption 4.1 is satisfied by a number of familiar moment selection criteria. Substituting Corollary 3.2 into Equation 3.6 shows that the FMSC converges to a function of M and constants only. Therefore, any almost surely continuous weights that can be written as a function of FMSC satisfy Assumption 4.1. Thus, we can use Corollary 2.1 to study the limit behavior of post-FMSC estimators. Moment selection criteria based on the J-test statistic also satisfy the conditions of Assumption 4.1. Under local mis-specification, the J-test statistic does not diverge, but has a non-central χ^2 limit distribution that can be expressed as a function of M and constants as follows.

Theorem 4.1 (Distribution of J-Statistic under Local Mis-Specification)**.** *Under the conditions of Corollary 2.1,*

$$J_n(S) = n\left[\Xi_S f_n(\widehat{\theta}_S)\right]' \widehat{\Omega}^{-1} \left[\Xi_S f_n(\widehat{\theta}_S)\right] \to_d \left(\Omega_S^{-1/2} M_S\right)' (I - P_S) \left(\Omega_S^{-1/2} M_S\right)$$

where $\widehat{\Omega}_S^{-1}$ *is a consistent estimator of* Ω_S^{-1}*,* P_S *is the projection matrix based on the identifying restrictions* $\Omega_S^{-1/2} F_S$*, and* $M_S = \Xi_S M$*.*

Thus, the downward J-test procedure, GMM-BIC, GMM-HQ, and GMM-AIC all satisfy Corollary 4.1. GMM-BIC and GMM-HQ, however, are not particularly interesting under local mis-specification. Intuitively, because they aim to select all valid moment conditions w.p.a.1, we would expect that under Assumption 2.1 they simply choose the full moment set in the limit. The following result states that this intuition is correct.

Theorem 4.2 (Behavior of Consistent Criteria under Local Mis-Specification)**.** *Consider a moment selection criterion of the form* $MSC(S) = J_n(S) - h(|S|)\kappa_n$*, where*

(a) h *is strictly increasing, and*

(b) $\kappa_n \to \infty$ *as* $n \to \infty$ *and* $\kappa_n = o(n)$*.*

Under the conditions of Corollary 2.1, $MSC(S)$ *selects the full moment set w.p.a.1.*

Because moment selection using the GMM-BIC or HQ leads to weights $\omega(M|S)$ with a degenerate distribution, these examples are not considered further below.

4.3. Valid Confidence Intervals

While Corollary 4.1 characterizes the limiting behavior of moment-average, and hence post-selection estimators, it does not immediately suggest a procedure for constructing confidence intervals. The limiting random variable $\Lambda(\tau)$ defined in Corollary 4.1 is a complicated function of the normal random vector M, the precise form of which depends on the weight function $\omega(\cdot|S)$. To surmount this difficulty, I adapt a suggestion from Claeskens and Hjort (2008) and approximate the behavior of moment average estimators by simulation. The result is a conservative procedure that provides asymptotically valid confidence intervals.

First, suppose that K_S, $\omega(\cdot|S)$, θ_0, Ω and τ are known. Then, by simulating from M, as defined in Theorem 2.2, the distribution of $\Lambda(\tau)$, defined in Corollary 4.1, can be approximated to arbitrary precision. To operationalize this procedure, substitute consistent estimators of K_S, θ_0, and Ω, e.g. those used to calculate FMSC. To estimate $\omega(\cdot|S)$, we first need to derive the limit distribution of $\widehat{\omega}(S)$, the data-based weight function specified by the user. As an example, consider the case of moment selection based on the FMSC. Here $\widehat{\omega}(S)$ is simply the indicator function

$$\widehat{\omega}(S) = \mathbf{1}\left\{\text{FMSC}_n(S) = \min_{S' \in \mathcal{A}} \text{FMSC}_n(S')\right\} \tag{4.2}$$

To estimate $\omega(\cdot|S)$ we require the limiting distribution of $\text{FMSC}_n(S)$. From 3.6, by Corollary 3.2, if $\widehat{\Omega} \to_p \Omega$, $\widehat{K}_S \to_p K_S$ and $\widehat{\theta} \to_p \theta_0$, $\text{FMSC}_n(S) \to_d \text{FMSC}(M|S)$ where

$$\text{FMSC}(M|S) = \nabla_\theta \mu(\theta_0)' K_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \Psi\left(MM' - \Omega\right)\Psi' \end{bmatrix} + \Omega \right\} \Xi' K_S' \nabla_\theta \mu(\theta_0) \tag{4.3}$$

Defining

$$\widehat{\text{FMSC}}(M|S) = \nabla_\theta \mu(\widehat{\theta})' \widehat{K}_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \widehat{\Psi}\left(MM' - \widehat{\Omega}\right)\widehat{\Psi}' \end{bmatrix} + \Omega \right\} \Xi_S' \widehat{K}_S' \nabla_\theta \mu(\widehat{\theta}) \tag{4.4}$$

yields the following estimator of $\omega(\cdot|S)$ for the case of FMSC moment selection

$$\widehat{\omega}(\cdot|S) = \mathbf{1}\left\{\widehat{\text{FMSC}}(\cdot|S) = \min_{S' \in \mathcal{A}} \widehat{\text{FMSC}}(\cdot|S')\right\} \tag{4.5}$$

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

Simulating from M, defined in Equation 2.7, requires estimates of Ω and τ. Recall that no consistent estimator of τ is available under local mis-specification; the estimator $\widehat{\tau}$ has a non-degenerate limit distribution (see Theorem 3.2). Thus, simulation from a $\mathcal{N}_{p+q}((0', \widehat{\tau}')', \widehat{\Omega})$ distribution may lead to erroneous results by failing to account for the uncertainty that enters through $\widehat{\tau}$. The solution is to use a two-stage procedure. First construct a $100(1-\delta)\%$ confidence region $T(\widehat{\tau})$ for τ using Theorem 3.2. Then simulate from the distribution of $\Lambda(\tau)$, defined in Corollary 4.1, for each $\tau \in T(\widehat{\tau})$. Taking the lower and upper bounds of the resulting intervals, centering and rescaling yields a conservative interval for $\widehat{\mu}$, as defined in defined in Equation 4.1. The precise algorithm is as follows.

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

1. For each $\tau \in T(\widehat{\tau})$

   (i) Generate $M_j(\tau) \sim \mathcal{N}_{p+q}\left((0', \tau')', \widehat{\Omega}\right)$, $j = 1, 2, \ldots, B$

   (ii) Set $\Lambda_j(\tau) = -\nabla_\theta \mu(\widehat{\theta})' \left[\sum_{S \in \mathcal{A}} \widehat{\omega}(M_j(\tau)|S) \widehat{K}_S \Xi_S\right] M_j(\tau)$, $j = 1, 2, \ldots, B$

   (iii) Using $\{\Lambda_j(\tau)\}_{j=1}^B$, calculate $\widehat{a}(\tau)$, $\widehat{b}(\tau)$ such that

   $$\mathbb{P}\left\{\widehat{a}(\tau) \leq \Lambda(\tau) \leq \widehat{b}(\tau)\right\} = 1 - \alpha$$

2. Define

   $$\begin{aligned} \widehat{a}_{min}(\widehat{\tau}) &= \min_{\tau \in T(\widehat{\tau})} \widehat{a}(\tau) \\ \widehat{b}_{max}(\widehat{\tau}) &= \max_{\tau \in T(\widehat{\tau})} \widehat{b}(\tau) \end{aligned}$$

3. The confidence interval for is μ given by

   $$\text{CI}_{sim} = \left[\widehat{\mu} - \frac{\widehat{b}_{max}(\widehat{\tau})}{\sqrt{n}}, \quad \widehat{\mu} - \frac{\widehat{a}_{min}(\widehat{\tau})}{\sqrt{n}}\right]$$

Theorem 4.3 (Simulation-based Confidence Interval for $\widehat{\mu}$)**.** *If*

(a) $\widehat{\Psi}$, $\widehat{\Omega}$, $\widehat{\theta}$ *and* $\widehat{K}_S$ *are consistent estimators of* Ψ, Ω, θ_0 *and* K_S*;*

(b) $\widehat{\omega}(M|S) = \omega(M|S) + o_p(1)$*;*

(c) $\widehat{\Delta}(\widehat{\tau}, \tau) = (\widehat{\tau} - \tau)' \left(\widehat{\Psi}\widehat{\Omega}\widehat{\Psi}'\right)^{-1} (\widehat{\tau} - \tau)$ *and*

(d) $T(\widehat{\tau}) = \left\{\tau \colon \Delta_n(\widehat{\tau}, \tau) \leq \chi^2_q(\delta)\right\}$ *where* $\chi^2_q(\delta)$ *denotes the* $1-\delta$ *quantile of a* χ^2 *distribution with* q *degrees of freedom*

Table 9: Coverage probabilities post-FMSC moment selection of a traditional 95% asymptotic confidence interval for the effect of x on y in Equation 3.17, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
0.1	0.91	0.87	0.88	0.91	0.93	0.93	0.93	0.93	0.93
0.2	0.90	0.79	0.72	0.82	0.90	0.93	0.92	0.93	0.93
0.3	0.90	0.76	0.58	0.64	0.80	0.90	0.92	0.93	0.93
0.4	0.89	0.75	0.50	0.47	0.64	0.80	0.88	0.91	0.92
0.5	0.89	0.74	0.45	0.36	0.50	0.67	0.79	0.87	0.91
0.6	0.89	0.74	0.43	0.30	0.38	0.54	0.68	0.78	0.85
0.7	0.90	0.74	0.41	0.24	0.31	0.44	0.57	0.68	0.78
0.8	0.89	0.74	0.41	0.22	0.25	0.36	0.48	0.59	0.70
0.9	0.91	0.74	0.41	0.20	0.21	0.31	0.41	0.52	0.61
1.0	0.90	0.75	0.40	0.18	0.19	0.25	0.35	0.45	0.53
1.1	0.90	0.76	0.40	0.17	0.17	0.23	0.32	0.39	0.47
1.2	0.91	0.76	0.41	0.17	0.15	0.20	0.27	0.34	0.42
1.3	0.92	0.77	0.41	0.16	0.15	0.19	0.24	0.31	0.39

$\gamma = Cov(w, x)$

then, the interval CI_{sim} defined in Algorithm 4.1 has asymptotic coverage probability no less than $1 - (\alpha + \delta)$ as $B, n \to \infty$.

To evaluate the performance of the procedure given in Algorithm 4.1, we revisit the simulation experiment described in Section 3.4, considering FMSC moment selection. The following results are based on 10,000 replications, each with a sample size of 500. Table 9 gives the empirical coverage probabilities of traditional 95% confidence intervals post-FMSC selection. These are far below the nominal level over the vast majority of the parameter space. Table 10 presents the empirical coverage of conservative 90% confidence intervals constructed according to Algorithm 4.1, with $B = 1000$.[2] The two-stage simulation procedure performs remarkably well, achieving a minimum coverage probability of 0.89 relative to its nominal level of 0.9. Moreover, a naïve one-step procedure that omits the first-stage and simply simulates from M based on $\widehat{\tau}$ performs surprisingly well; see Table 11. While the empirical coverage probabilities of the one-step procedure are generally lower than the nominal level of 0.95, they represent a substantial improvement over the traditional intervals given in Table 9, with a worst-case coverage of 0.72 compared to 0.15. This suggests that the one-step intervals might be used as a rough but useful approximation to the fully robust but more computationally intensive intervals constructed according to Algorithm 4.1.

[2]Because this simulation is computationally intensive, I use a reduced grid of parameter values.

Table 10: Coverage probabilities of two-step, conservative 90% intervals for the effect of x on y in Equation 3.17, post-FMSC moment selection. Intervals are calculated using Algorithm 4.1 with $B = 1000$, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. As above, simulations are generated from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

		$\rho = Cov(w,u)$				
$N = 500$		0	0.1	0.2	0.3	0.4
$\gamma = Cov(w,x)$	0.0	0.92	0.93	0.93	0.93	0.94
	0.2	0.95	0.91	0.93	0.95	0.97
	0.4	0.95	0.95	0.90	0.93	0.97
	0.6	0.95	0.95	0.92	0.90	0.92
	0.8	0.94	0.95	0.96	0.90	0.89
	1.0	0.94	0.94	0.96	0.93	0.90
	1.2	0.94	0.94	0.96	0.95	0.92

Table 11: Coverage probabilities of corrected one-step, 95% intervals for the effect of x on y in Equation 3.17, post-FMSC moment selection. Intervals are calculated using Step 1 of Algorithm 4.1, fixing $\tau = \widehat{\tau}$, with $B = 1000$, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. As above, simulations are generated from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

		$\rho = Cov(w,u)$								
$N = 500$		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w,x)$	0.0	0.93	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.94
	0.1	0.93	0.91	0.91	0.92	0.92	0.92	0.93	0.94	0.95
	0.2	0.94	0.91	0.86	0.87	0.92	0.93	0.94	0.95	0.96
	0.3	0.95	0.94	0.87	0.81	0.85	0.91	0.94	0.96	0.96
	0.4	0.95	0.95	0.91	0.82	0.77	0.84	0.90	0.94	0.95
	0.5	0.95	0.95	0.93	0.86	0.76	0.76	0.82	0.88	0.92
	0.6	0.94	0.94	0.94	0.90	0.80	0.74	0.75	0.81	0.87
	0.7	0.94	0.94	0.95	0.93	0.85	0.74	0.73	0.75	0.81
	0.8	0.94	0.94	0.95	0.94	0.88	0.79	0.73	0.73	0.76
	0.9	0.95	0.94	0.94	0.94	0.91	0.83	0.76	0.72	0.73
	1.0	0.95	0.94	0.94	0.94	0.92	0.86	0.78	0.73	0.73
	1.1	0.95	0.94	0.94	0.95	0.94	0.89	0.81	0.76	0.73
	1.2	0.95	0.94	0.94	0.95	0.94	0.90	0.85	0.79	0.75
	1.3	0.95	0.94	0.94	0.95	0.95	0.92	0.87	0.81	0.78

4.4. Moment Averaging

The moment average estimators of the previous section were derived primarily to provide valid confidence intervals post-moment selection, but in fact allow us to carry out inference for a wider class of estimators. Viewed as a special case of Equation 4.1, moment selection is in fact a fairly crude procedure, giving full weight to the minimizer of the criterion no matter how close its nearest competitor lies. Under moment selection, when competing moment sets have similar criterion values in the population, random variation in the sample will be magnified in the selected estimator. Thus, it may be possible to achieve better performance by using smooth weights rather than discrete selection. In this section, I briefly examine a proposal based on exponential weighting.

In the context of maximum likelihood estimation, Buckland et al. (1997) suggest averaging the estimators resulting from a number of competing models using weights of the form

$$w_k = \frac{\exp(-I_k/2)}{\sum_{i=1}^{K} \exp(-I_i/2)} \tag{4.6}$$

where I_k is an information criterion evaluated for model k, and $i \in \{1, 2, \ldots, K\}$ indexes the set of candidate models. This expression, constructed by an analogy with Bayesian model averaging, gives more weight to models with lower values of the information criterion but non-zero weight to all models. Applying this idea to the moment selection criteria given above, consider

$$\widehat{\omega}_{BIC}(S) = \exp\left\{-\frac{\kappa}{2}\text{GMM-BIC}(S)\right\} \Big/ \sum_{S' \in \mathcal{A}} \exp\left\{-\frac{\kappa}{2}\text{GMM-BIC}(S')\right\} \tag{4.7}$$

$$\widehat{\omega}_{AIC}(S) = \exp\left\{-\frac{\kappa}{2}\text{GMM-AIC}(S)\right\} \Big/ \sum_{S' \in \mathcal{A}} \exp\left\{-\frac{\kappa}{2}\text{GMM-AIC}(S')\right\} \tag{4.8}$$

$$\widehat{\omega}_{HQ}(S) = \exp\left\{-\frac{\kappa}{2}\text{GMM-HQ}(S)\right\} \Big/ \sum_{S' \in \mathcal{A}} \exp\left\{-\frac{\kappa}{2}\text{GMM-HQ}(S')\right\} \tag{4.9}$$

$$\widehat{\omega}_{FMSC}(S) = \exp\left\{-\frac{\kappa}{2}\text{FMSC}(S)\right\} \Big/ \sum_{S' \in \mathcal{A}} \exp\left\{-\frac{\kappa}{2}\text{FMSC}(S')\right\} \tag{4.10}$$

The parameter κ varies the uniformity of the weighting. As $\kappa \to 0$ the weights become more uniform; as $\kappa \to \infty$ they approach the moment selection procedure given by minimizing the corresponding criterion.

Table 12 compares moment averaging based on Equations 4.7–4.10 to the corresponding moment selection procedures using the simulation experiment described in Section 3.4. Calculations are based on 10,000 replications, each with a sample size of 500. For FMSC averaging $\kappa = 1/100$ to account for the fact that the FMSC is generally more variable than criteria based on the J-test. Weights for GMM-BIC, HQ, and AIC averaging set $\kappa = 1$. Both in terms of average and worst-case RMSE, moment selection is inferior to moment averaging. The only exception is worst-case RMSE for the FMSC. Moreover, as we see from Tables 13–16, which compare the averaging and selection procedures at each point on the simulation grid, this improvement is nearly uniform. If our goal is estimators with low RMSE, moment averaging may be preferable to moment selection.

Table 12: Average and worst-case RMSE of the moment averaging procedures given in Equations 4.7–4.10 and their moment selection counterparts, with $\kappa = 1/100$ for FMSC averaging and $\kappa = 1$ for all other averaging procedures. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications at each combination of parameter values from Table 1 and a sample size of 500.

Average RMSE	Averaging	Selection
FMSC	0.24	0.26
GMM-BIC	0.26	0.29
GMM-HQ	0.26	0.29
GMM-AIC	0.26	0.28
Worst-Case RMSE	Averaging	Selection
FMSC	0.36	0.33
GMM-BIC	0.41	0.47
GMM-HQ	0.36	0.39
GMM-AIC	0.33	0.35

Table 13: Difference in RMSE between GMM-BIC moment averaging with $\kappa = 1$ and GMM-BIC moment selection, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that averaging gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02
0.1	-0.01	-0.01	-0.03	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02
0.2	0.00	-0.02	-0.04	-0.04	-0.02	-0.02	-0.02	-0.02	-0.02
0.3	-0.01	-0.02	-0.04	-0.05	-0.05	-0.04	-0.03	-0.02	-0.03
0.4	-0.01	-0.02	-0.03	-0.04	-0.05	-0.05	-0.04	-0.04	-0.03
0.5	-0.01	-0.01	-0.03	-0.04	-0.05	-0.06	-0.05	-0.05	-0.05
0.6	-0.01	-0.01	-0.02	-0.03	-0.04	-0.05	-0.06	-0.06	-0.06
0.7	-0.01	-0.01	-0.02	-0.04	-0.04	-0.05	-0.05	-0.06	-0.06
0.8	-0.02	-0.02	-0.02	-0.03	-0.04	-0.04	-0.05	-0.05	-0.06
0.9	-0.01	-0.02	-0.01	-0.03	-0.03	-0.04	-0.05	-0.05	-0.06
1.0	-0.02	-0.02	-0.03	-0.02	-0.03	-0.04	-0.04	-0.05	-0.05
1.1	-0.01	-0.01	-0.01	-0.02	-0.03	-0.03	-0.04	-0.04	-0.05
1.2	-0.02	-0.01	-0.02	-0.02	-0.03	-0.03	-0.04	-0.04	-0.04
1.3	-0.01	-0.02	-0.02	-0.02	-0.03	-0.03	-0.03	-0.04	-0.04

$\gamma = Cov(w, x)$

Table 14: Difference in RMSE between GMM-HQ moment averaging with $\kappa = 1$ and GMM-HQ moment selection, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that averaging gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.1	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.01	-0.02
0.2	-0.01	-0.02	-0.03	-0.02	-0.01	0.00	-0.01	0.00	-0.01
0.3	-0.01	-0.03	-0.04	-0.04	-0.03	-0.01	0.00	0.00	0.00
0.4	-0.02	-0.02	-0.03	-0.04	-0.04	-0.03	-0.02	-0.01	-0.01
0.5	-0.02	-0.03	-0.04	-0.04	-0.04	-0.04	-0.03	-0.02	-0.02
0.6	-0.02	-0.02	-0.03	-0.04	-0.04	-0.05	-0.04	-0.04	-0.03
0.7	-0.02	-0.03	-0.03	-0.04	-0.05	-0.05	-0.05	-0.05	-0.04
0.8	-0.02	-0.03	-0.03	-0.04	-0.04	-0.05	-0.05	-0.05	-0.05
0.9	-0.03	-0.02	-0.03	-0.04	-0.04	-0.04	-0.05	-0.05	-0.05
1.0	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04	-0.05	-0.05	-0.05
1.1	-0.03	-0.03	-0.04	-0.04	-0.04	-0.04	-0.05	-0.05	-0.05
1.2	-0.03	-0.02	-0.03	-0.03	-0.04	-0.04	-0.04	-0.05	-0.05
1.3	-0.05	-0.02	-0.03	-0.03	-0.04	-0.04	-0.04	-0.04	-0.05

$\gamma = Cov(w, x)$

Table 15: Difference in RMSE between GMM-AIC moment averaging with $\kappa = 1$ and GMM-AIC moment selection, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that averaging gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00
0.1	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.01	-0.01
0.2	-0.02	-0.02	-0.02	0.00	0.01	0.01	0.00	0.00	0.00
0.3	-0.03	-0.03	-0.03	-0.01	0.00	0.01	0.01	0.01	0.00
0.4	-0.03	-0.03	-0.03	-0.03	-0.01	0.00	0.00	0.01	0.01
0.5	-0.03	-0.03	-0.03	-0.03	-0.02	-0.01	-0.01	0.00	0.00
0.6	-0.03	-0.03	-0.03	-0.04	-0.03	-0.02	-0.02	-0.01	-0.01
0.7	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.02	-0.02
0.8	-0.03	-0.03	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.02
0.9	-0.04	-0.03	-0.03	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03
1.0	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03
1.1	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
1.2	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
1.3	-0.05	-0.04	-0.04	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04

$\gamma = Cov(w, x)$

Table 16: Difference in RMSE between FMSC moment averaging with $\kappa = 1/100$ and FMSC moment selection, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that averaging gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	0.00	0.00	0.01	0.02	0.02	0.02	0.02	0.02
0.1	-0.01	-0.01	-0.01	0.03	0.05	0.06	0.06	0.06	0.06
0.2	-0.02	-0.03	-0.04	0.00	0.05	0.08	0.08	0.07	0.06
0.3	-0.03	-0.04	-0.05	-0.04	0.01	0.05	0.08	0.09	0.08
0.4	-0.04	-0.04	-0.05	-0.06	-0.03	0.01	0.04	0.07	0.08
0.5	-0.04	-0.04	-0.05	-0.06	-0.05	-0.03	0.00	0.03	0.06
0.6	-0.04	-0.04	-0.05	-0.06	-0.06	-0.05	-0.03	0.00	0.03
0.7	-0.04	-0.04	-0.05	-0.06	-0.06	-0.06	-0.05	-0.03	0.00
0.8	-0.04	-0.04	-0.05	-0.05	-0.06	-0.06	-0.05	-0.05	-0.02
0.9	-0.04	-0.04	-0.04	-0.05	-0.06	-0.06	-0.06	-0.05	-0.04
1.0	-0.04	-0.04	-0.04	-0.05	-0.05	-0.06	-0.06	-0.05	-0.05
1.1	-0.04	-0.04	-0.04	-0.04	-0.05	-0.06	-0.06	-0.06	-0.06
1.2	-0.04	-0.03	-0.03	-0.04	-0.05	-0.05	-0.06	-0.06	-0.06
1.3	-0.04	-0.04	-0.04	-0.04	-0.04	-0.05	-0.05	-0.06	-0.06

$\gamma = Cov(w, x)$

5. Empirical Example: Geography or Institutions?

Carstensen and Gundlach (2006) address a controversial question from the development literature: does geography directly effect income after controlling for institutions? A number of well-known studies find little or no direct effect of geographic endowments. Acemoglu et al. (2001), for example, find that countries nearer to the equator do not have lower incomes after controlling for institutions. Rodrik et al. (2004) report that geographic variables have only small direct effects on income, affecting development mainly through their influence on institutions. Similarly, Easterly and Levine (2003) find no effect of "tropics, germs and crops" except through institutions. Sachs (2003) responds directly to these three papers by showing that malaria transmission, a variable largely driven by ecological conditions, directly influences the level of per capita income, even after controlling for institutions. Because malaria transmission is very likely endogenous, Sachs uses a measure of "malaria ecology," constructed to be exogenous both to present economic conditions and public health interventions, as an instrument. Carstensen and Gundlach (2006) address the robustness of Sachs's results using the following baseline regression for a sample of 45 countries:

$$\ln gdpc_i = \beta_1 + \beta_2 \cdot institutions_i + \beta_3 \cdot malaria_i + \epsilon_i \tag{5.1}$$

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

section, using data kindly supplied by the authors, I expand on the instrument selection exercise given in Table 2 of Carstensen and Gundlach (2006) using the FMSC and corrected confidence intervals described above. I consider two questions. First, based on the FMSC methodology, which instruments should we choose to produce the best estimate of β_3, the effect of malaria transmission on per capita income? Second, after correcting confidence intervals for instrument selection, do we still find evidence of large and negative effects of malaria transmission on income? All results given here are calculated by 2SLS using the formulas from Section 3.3 and the variables described in Table 17. In keeping with Table 2 of Carstensen and Gundlach (2006), I use ln*gdpc* as the dependent variable and *rule* and *malfal* as measures of institutions and malaria transmission throughout this section.

To apply the FMSC to the present example, we need a minimum of two valid instruments besides the constant term. Based on the arguments given in Acemoglu et al. (2001), Carstensen and Gundlach (2006) and Sachs (2003), I proceed under the assumption that ln*mort* and *maleco*, measures of early settler mortality and malaria ecology, are exogenous. Rather than selecting over every possible subset of instruments, I consider a number of instrument blocks defined in Carstensen and Gundlach (2006). The baseline block contains ln*mort*, *maleco* and a constant; the climate block contains *frost*, *humid*, and *latitude*; the Europe block contains *eurfrac* and *engfrac*; and the openness block contains *coast* and *trade*. Full descriptions of these variables appear in Table 17. Table 18 gives 2SLS results and traditional 95% confidence intervals for all instrument sets considered here.

Table 19 presents FMSC results for instrument sets 1–8 as defined in Table 18. The left panel takes the effect of *malfal*, a measure of malaria transmission, as the target parameter while the right uses the effect of *rule*, a measure of institutions. Results are sorted in decreasing order of FMSC, with the selected instrument set and corresponding estimate at the bottom. In each case, the FMSC selects instrument set 8: the full instrument set containing the baseline, climate, Europe and openness blocks. The FMSC rankings, however, differ depending on the target parameter. For example, when the target is *rule* instrument sets 8 and 5 are virtually identical in terms of FMSC: 0.26 versus 0.23. In Table 2 of their paper, Carstensen and Gundlach (2006) report GMM-BIC and HQ results for selection over instrument sets 2–4 and 8 that also favor instrument set 8. However, the authors do not consider instrument sets 5–7. Although the FMSC also selects instrument set 8, the FMSC values of instrument set 5 are small enough to suggest that including the openness block does little to reduce MSE.

The bottom two panels of Table 19 present a number of alternative 95% confidence intervals for the effects of *malfal* and *rule*, respectively. The first row gives the traditional asymptotic confidence interval from Table 18, while the following three give simulation-based intervals accounting for the effects of instrument selection. I do not present intervals for the conservative procedure given in Algorithm 4.1 because the results in this example are so insensitive to the value of τ that the minimization and maximization problems given in Step 2 of the Algorithm are badly behaved. To illustrate this, I instead present intervals that use the same simulation procedure as Algorithm 4.1 but treat τ as fixed. I consider four possible values of the bias parameter. When $\tau = \widehat{\tau}$, we have the one-step corrected interval considered in Table 11. When $\tau = 0$, we have an interval that assumes all instruments are valid. The remaining two values $\widehat{\tau}_{min}$ and $\widehat{\tau}_{max}$ correspond to the lower and upper bounds of *elementwise* 95% confidence intervals for τ based on the distributional result given in

Table 17: Description of Variables

Name	Description	Type
ln*gdpc*	Real GDP/capita at PPP, 1995 International Dollars	Dependent Variable
rule	Institutional quality (Average Governance Indicator)	Regressor
malfal	Proportion of population at risk of malaria transmission, 1994	Regressor
ln*mort*	Log settler mortality (per 1000 settlers), early 19th century	Baseline Instrument
maleco	Index of stability of malaria transmission	Baseline Instrument
frost	Proportion of land receiving at least 5 days of frost in winter	Climate Instrument
humid	Highest temperature ($^{\circ}C$) in month with highest average afternoon humidity	Climate Instrument
latitude	Distance from equator (absolute value of latitude in degrees)	Climate Instrument
eurfrac	Porportion of population that speaks a major Western European Language	Europe Instrument
engfrac	Proportion of population that speaks English	Europe Instrument
coast	Proportion of land area within 100km of sea coast	Openness Instrument
trade	Log Frankel-Romer predicted trade share, using population and geography	Openness Instrument

Table 18: 2SLS Results for all Instrument Sets

	1		2		3		4		5		6	
	rule	*malfal*	*rule*	*malfal*	*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	0.93	-1.02	0.86	-0.98
SE	0.18	0.30	0.16	0.29	0.16	0.29	0.16	0.27	0.15	0.26	0.14	0.27
lower	0.53	-1.65	0.65	-1.48	0.49	-1.67	0.55	-1.69	0.63	-1.54	0.59	-1.53
upper	1.25	-0.43	1.30	-0.32	1.13	-0.51	1.18	-0.59	1.22	-0.49	1.14	-0.43
	Baseline		Baseline		Baseline		Baseline		Baseline		Baseline	
			Climate						Climate		Climate	
					Openness						Openness	
							Europe		Europe			

	7		8		9		10		11		12	
	rule	*malfal*	*rule*	*malfal*	*rule*	*malfal*	*rule*	*malfal*	*rule*	*malfal*	*rule*	*malfal*
coeff.	0.81	-1.16	0.84	-1.08	0.93	-0.93	1.02	-0.85	1.02	-0.86	0.88	-1.00
SE	0.15	0.27	0.13	0.25	0.16	0.23	0.15	0.27	0.15	0.23	0.12	0.21
lower	0.51	-1.70	0.57	-1.58	0.61	-1.39	0.71	-1.39	0.72	-1.32	0.63	-1.42
upper	1.11	-0.62	1.10	-0.58	1.26	-0.46	1.33	-0.30	1.32	-0.40	1.12	-0.57
	Baseline		Baseline		Baseline		Baseline		Baseline		Baseline	
			Climate								Climate	
	Openness		Openness								Openness	
	Europe		Europe								Europe	
					$malfal^2$				$malfal^2$		$malfal^2$	
							$rule^2$		$rule^2$		$rule^2$	

Theorem 3.1. These result in a region with greater than 95% coverage for τ considered jointly. We see that the corrected 95% intervals for the effect of *malfal* are extremely similar regardless of the value of τ used in the simulation. The same is true for *rule*. There is no evidence that accounting for the effects of instrument selection should change our conclusions about the sign or significance of either *malfal* or *rule*.

Table 19: FMSC values and corrected confidence intervals for selection over instrument sets 1–8. The left panel gives results when the coefficient on *malfal* is the target parameter; the right panel gives results when the coefficient on *rule* is the target parameter.

μ =*malfal*	FMSC	$\widehat{\mu}$
Valid (1)	3.03	-1.04
Climate (2)	2.67	-0.90
Openness (3)	2.31	-1.09
Europe (4)	1.83	-1.14
Openness, Europe (7)	1.72	-1.16
Climate, Openness (6)	1.65	-0.98
Climate, Europe (5)	0.71	-1.02
Full (8)	0.53	-1.08

μ =*rule*	FMSC	$\widehat{\mu}$
Valid (1)	1.27	0.89
Openness (3)	1.23	0.81
Climate (2)	0.92	0.97
Openness, Europe (7)	0.77	0.81
Europe (4)	0.55	0.86
Climate, Openness (6)	0.43	0.86
Climate, Europe (5)	0.26	0.93
Full (8)	0.23	0.84

μ =*malfal*	lower	upper
Traditional	-1.58	-0.58
$\tau = \widehat{\tau}$	-1.54	-0.61
$\tau = 0$	-1.53	-0.64
$\tau = \widehat{\tau}_{max}$	-1.51	-0.55
$\tau = \widehat{\tau}_{min}$	-1.61	-0.58

μ =*rule*	lower	upper
Traditional	0.57	1.10
$\tau = \widehat{\tau}$	0.55	1.13
$\tau = 0$	0.55	1.12
$\tau = \widehat{\tau}_{max}$	0.55	1.17
$\tau = \widehat{\tau}_{min}$	0.49	1.15

The FMSC is designed to include invalid instruments when doing so will reduce AMSE. Table 20 considers adding two almost certainly invalid instruments to the baseline instrument set: $rule^2$ and $malfal^2$. Because they are constructed from the endogenous regressors, these instruments are likely to be highly relevant. Unless the effect of institutions and malaria transmission on GDP per capita is exactly linear, however, they are invalid. When the target is *malfal*, we see that the FMSC selects an instrument set including $malfal^2$ and the baseline instruments. Notice that the FMSC is negative in this case. Although it provides an asymptotically unbiased estimator of AMSE, the FMSC may be negative because it subtracts $\widehat{\Psi}\widehat{\Omega}\widehat{\Psi}'$ from $\widehat{\tau}\widehat{\tau}'$ when estimating the squared bias. When the target is *rule*, the FMSC chooses the full instrument set, including the baseline instruments along with $rule^2$ and $malfal^2$. While these instruments are likely invalid, they are extremely strong. The FMSC chooses to include them because its estimate of the bias they induce is small compared to the reduction in variance they provide. Table 21 further expands the instrument sets under consideration to include 1–4 and 9–12. In this case, the FMSC chooses instrument set 12 for both target parameters. However, we see from the FMSC rankings that most of the reduction in MSE achieved by instrument set 12 comes from the inclusion of the squared endogenous regressors in the instrument set. Turning our attention to the confidence intervals in Tables 20 and 21,

we again see that the simulation-based intervals are extremely insensitive to the value of τ used. Once again, the sign and significance of *malfal* and *rule* is not sensitive to the effects of instrument selection. These results lend further support to the view of Carstensen and Gundlach (2006) and Sachs (2003) that malaria transmission has a direct effect on economic development.

Table 20: FMSC values and corrected confidence intervals for selection over instrument sets 1 and 9–11. The left panel gives results when the coefficient on *malfal* is the target parameter; the right panel gives results when the coefficient on *rule* is the target parameter.

μ =*malfal*	FMSC	$\widehat{\mu}$
Valid (1)	3.03	-1.04
$rule^2$ (10)	2.05	-0.84
Full (11)	-0.20	-0.85
$malfal^2$ (9)	-0.41	-0.92

μ =*malfal*	lower	upper
Traditional	-1.39	-0.46
$\tau = \widehat{\tau}$	-1.49	-0.38
$\tau = 0$	-1.46	-0.38
$\tau = \widehat{\tau}_{max}$	-1.51	-0.38
$\tau = \widehat{\tau}_{min}$	-1.49	-0.38

μ =*rule*	FMSC	$\widehat{\mu}$
Valid (1)	1.27	0.89
$rule^2$ (10)	0.28	1.02
$malfal^2$ (9)	0.18	0.93
Full (11)	-0.06	1.02

μ =*rule*	lower	upper
Traditional	0.72	1.32
$\tau = \widehat{\tau}$	0.68	1.36
$\tau = 0$	0.71	1.32
$\tau = \widehat{\tau}_{max}$	0.66	1.37
$\tau = \widehat{\tau}_{min}$	0.71	1.35

6. Conclusion

This paper has introduced the FMSC, a proposal to choose moment conditions based on the quality of the estimates they provide. The criterion performs well in simulations, and the framework used to derive it allows us to construct confidence intervals that properly account for the effects of moment selection on inference. While I focus here on an application to instrument selection for cross-section data, 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 the stronger assumption 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. Similarly, in a macro model, measurement error could be present in variables entering the intratemporal Euler equation but not the intertemporal Euler equation, as considered by Eichenbaum et al. (1988). In this case we could use the FMSC to decide whether to include the moment conditions arising from the intra-Euler. Because the FMSC uses only first-order asymptotics, a possible extension of this work would be to consider refinements based on higher-order expansions. Another possibility would be to derive a version of the FMSC for generalized empirical likelihood (GEL) estimators. While GMM and GEL are first-order equivalent, GEL often gives finite-sample performance (Newey and Smith, 2004), and may thus improve the performance of the moment selection procedure.

Table 21: FMSC values and corrected confidence intervals for selection over instrument sets 1–4 and 9–12. The left panel gives results when the coefficient on *malfal* is the target parameter; the right panel gives results when the coefficient on *rule* is the target parameter.

μ =*malfal*	FMSC	$\widehat{\mu}$
Valid (1)	3.03	-1.04
Climate (2)	2.85	-0.90
Openness (3)	2.51	-1.09
Europe (4)	1.94	-1.14
$rule^2$ (10)	1.88	-0.84
$malfal^2$, $rule^2$ (11)	0.06	-0.85
$malfal^2$ (9)	-0.20	-0.92
Full (12)	-1.38	-1.00

μ =*rule*	FMSC	$\widehat{\mu}$
Valid (1)	1.27	0.89
Openness (3)	1.26	0.81
Climate (2)	0.95	0.97
Europe (4)	0.58	0.86
$rule^2$ (10)	0.25	1.02
$malfal^2$ (9)	0.15	0.93
$malfal^2$, $rule^2$ (11)	-0.03	1.02
Full (12)	-0.61	0.88

μ =*malfal*	lower	upper
Traditional	-1.42	-0.57
$\tau = \widehat{\tau}$	-1.51	-0.51
$\tau = 0$	-1.48	-0.52
$\tau = \widehat{\tau}_{max}$	-1.50	-0.50
$\tau = \widehat{\tau}_{min}$	-1.50	-0.49

μ =*rule*	lower	upper
Traditional	0.63	1.12
$\tau = \widehat{\tau}$	0.57	1.17
$\tau = 0$	0.60	1.15
$\tau = \widehat{\tau}_{max}$	0.55	1.17
$\tau = \widehat{\tau}_{min}$	0.59	1.18

References

Acemoglu, D., Johnson, S., Robinson, J. A., 2001. The colonial origins of comparative development: An empirical investigation. American Economic Review 91 (5), 1369–1401.

Andrews, D. W. K., May 1999. Consistent moment selection procedures for generalized methods of moments estimation. Econometrica 67 (3), 543–564.

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

Buckland, S. T., Burnham, K. P., Augustin, N. H., 1997. Model selection: An integral part of inference. Biometrics 53 (2), 603–618.

Carstensen, K., Gundlach, E., 2006. The primacy of institutions reconsidered: Direct income effects of malaria prevelance. World Bank Economic Review 20 (3), 309–339.

Claeskens, G., Hjort, N. L., 2003. The focused information criterion. Journal of the American Statistical Association 98 (464), 900–945.

Claeskens, G., Hjort, N. L., 2008. Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge.

Conley, T. G., Hansen, C. B., Rossi, P. E., 2010. Plausibly exogenous. Forthcoming, Review of Economics and Statistics.

Davidson, J., 1994. Stochastic Limit Theory. Advanced Texts in Econometrics. Oxford.

Demetrescu, M., Hassler, U., Kuzin, V., 2011. Pitfalls of post-model-selection testing: Experimental quantification. Empirical Economics 40, 359–372.

Donald, S. G., Imbens, G. W., Newey, W. K., 2009. Choosing instrumental variables in conditional moment restriction models. Journal of Econometrics 152, 28–36.

Donald, S. G., Newey, W. K., September 2001. Choosing the number of instruments. Econometrica 69 (5), 1161–1191.

Easterly, W., Levine, R., 2003. Tropics, germs, and crops: how endowments influence economic development. Journal of Monetary Economics 50, 3–39.

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

Hall, A. R., 2005. Generalized Method of Moments. Advanced Texts in Econometrics. Oxford.

Hall, A. R., Peixe, F. P., 2003. A consistent method for the selection of relevant instruments in linear models. Econometric Reviews 22, 269–288.

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

Hong, H., Preston, B., Shum, M., 2003. Generalized empirical likelihood-based model selection for moment condition models. Econometric Theory 19, 923–943.

Jana, K., 2005. Canonical correlations and instrument selection in econometrics. Ph.D. thesis, North Carolina State University.

Kabaila, P., 1998. Valid confidence intervals in regressions after variable selection. Econometric Theory 14, 463–482.

Kabaila, P., Leeb, H., 2006. On the large-sample minimal coverage probability of confidence intervals after model selection. Journal of the American Statistical Association 101 (474), 819–829.

Kraay, A., 2010. Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach. Forthcoming, Journal of Applied Econometrics.

Kuersteiner, G., Okui, R., March 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., 2009. Model selection. In: Handbook of Financial Time Series. Springer.

Liao, Z., November 2010. Adaptive GMM shrinkage estimation with consistent moment selection, Working Paper.

Newey, W. K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. Vol. IV. Elsevier Science, Ch. 36, pp. 2111–2245.

Newey, W. K., Smith, R. J., 2004. Higher order properties of gmm and generalized empirical likelihood. Econometrica 72 (1), 219–255.

Phillips, P. C. B., 1980. The exact distribution of instrumental variables estimators in an equation containing $n + 1$ endogenous variables. Econometrica 48 (4), 861–878.

Pötscher, B. M., 1991. Effects of model selection on inference. Econometric Theory 7, 163–185.

Rodrik, D., Subramanian, A., Trebbi, F., 2004. Institutions rule: The primacy of institutions over geography and integration in economic development. Journal of Economic Growth 9, 131–165.

Sachs, J. D., February 2003. Institutions don't rule: Direct effects of geography on per capita income, NBER Working Paper No. 9490.

Xiao, Z., 2010. The weighted method of moments approach for moment condition models. Economics Letters 107, 183–186.

Appendix A. Proofs

Proof of Theorem 2.1. Essentially identical to the proof of Newey and McFadden (1994) Theorem 2.6. □

Proof of Theorem 2.2. C.f. the proof of Newey and McFadden (1994) Theorem 3.1. The only difference is that the proof here involves a normal vector with non-zero mean. □

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

$$\begin{aligned}\widehat{\tau} &= \sqrt{n}h_n\left(\widehat{\theta}_{valid}\right) = \sqrt{n}h_n(\theta_0) + H\sqrt{n}\left(\widehat{\theta}_{valid} - \theta_0\right) + o_p(1) \\ &= -HK_v\sqrt{n}f_n(\theta_0) + \mathbf{I}_q\sqrt{n}h_n(\theta_0) + o_p(1) \\ &= \Psi\sqrt{n}f_n(\theta_0) + o_p(1) \to_d \Psi M\end{aligned}$$

Multiplying through,

$$\mathbb{E}\left[\Psi M\right] = \Psi\mathbb{E}[M] = \left[\begin{array}{cc} -HK_v & \mathbf{I}_q \end{array}\right]\left[\begin{array}{c} 0 \\ \tau \end{array}\right] = \tau$$

and $Var\left[\Psi M\right] = \Psi\Omega\Psi'$. □

Proof of Corollary 3.2. By Theorem 3.1 and the Continuous Mapping Theorem,

$$\widehat{\tau}\widehat{\tau}' \to_d \Psi MM'\Psi'.$$

Since

$$\mathbb{V}[\Psi M] = \mathbb{E}[\Psi MM'\Psi'] - \mathbb{E}[\Psi M]\mathbb{E}[\Psi M]' = \mathbb{E}[\Psi MM'\Psi'] - \tau\tau'$$

we have

$$\mathbb{E}[\Psi MM'\Psi'] = \mathbb{V}[\Psi M] + \tau\tau = \Psi\Omega\Psi' + \tau\tau'.$$

□

Proof of Corollary 4.1. Because the weights sum to one

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

By Corollary 3.1,

$$\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right) \to_d -\nabla_\theta\mu(\theta_0)'K_SM_S.$$

By assumption $\widehat{\omega}(S) \to_d \omega(M|S)$ where $\omega(M|S)$ is a function of M and constants only. Hence $\widehat{\omega}(\cdot)$ and $\sqrt{n}\left(\widehat{\mu}(\cdot) - \mu_0\right)$ convergence jointly in distribution to their respective functions of M, for $S \in \mathcal{A}$. Therefore, applying the Continuous Mapping Theorem,

$$\sqrt{n}\left(\widehat{\mu} - \mu_0\right) \to_d -\nabla_\theta\mu(\theta_0)'\left[\sum_{S\in\mathcal{A}}\omega(M|S)K_S\Xi_S\right]M$$

since the weight functions are almost surely continuous. □

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

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

Since

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

we have

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

By Assumption 2.2 (viii), $\sqrt{n}\left[\Xi_S f_n(\theta_0)\right] \to_d M_S$. Thus, for estimation using the efficient weighting matrix

$$\widehat{\Omega}_S^{-1/2}\sqrt{n}\left[\Xi_S f_n(\theta_0)\right] \to_d \left[I - P_S\right]\Omega_S^{-1/2} M_S$$

where $\widehat{\Omega}_S^{-1/2}$ is a consistent estimator of $\Omega_S^{-1/2}$ and P_S is the projection matrix based on $\Omega_S^{-1/2} F_S$, the identifying restrictions.[3] Combining and rearranging,

$$J_n(S) = n\left[\Xi_S f_n(\widehat{\theta}_S)\right]' \widehat{\Omega}^{-1}\left[\Xi_S f_n(\widehat{\theta}_S)\right] \to_d \left(\Omega_S^{-1/2} M_S\right)'\left(I - P_S\right)\left(\Omega_S^{-1/2} M_S\right).$$

□

Proof of Theorem 4.2. Let S_1 and S_2 be arbitrary moment sets in $\mathcal{A}$, i.e. two subsets of the full moment set $S_{Full} = \{1, 2, \ldots, q\}$, and let $|S|$ denote the cardinality of S. By Theorem 4.1, $J_n(S) = O_p(1)$, $S \in \mathcal{A}$, thus

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

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

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

The result follows because $|S_{full}| > |S|$ for any $S \neq S_{full}$. □

Proof of Theorem 4.3. We have

$$\begin{aligned} \mathbb{P}\left\{\mu_{true} \in \mathrm{CI}_{sim}\right\} &= \mathbb{P}\left\{\widehat{\mu} - \widehat{b}_0(\widehat{\tau})/\sqrt{n} \leq \mu_{true} \leq \widehat{\mu} - \widehat{a}_0(\widehat{\tau})/\sqrt{n}\right\} \\ &= \mathbb{P}\left\{\widehat{a}_0(\widehat{\tau}) \leq \sqrt{n}\left(\widehat{\mu} - \mu_{true}\right) \leq \widehat{b}_0(\widehat{\tau})\right\} \end{aligned}$$

[3] See Hall (2005), Chapter 3.

By Theorem 3.1 and Corollary 4.1,

$$\begin{bmatrix} \sqrt{n}(\hat{\mu} - \mu_{true}) \\ \widehat{\tau} \end{bmatrix} \rightarrow_d \begin{bmatrix} \Lambda(\tau) \\ \Psi M \end{bmatrix}.$$

Thus

$$\begin{bmatrix} \sqrt{n}(\hat{\mu} - \mu_{true}) \\ \widehat{\Delta}(\widehat{\tau}, \tau) \\ \widehat{a}_{min}(\widehat{\tau}) \\ \widehat{b}_{max}(\widehat{\tau}) \end{bmatrix} \rightarrow_d \begin{bmatrix} \Lambda(\tau) \\ \Delta(\Psi M, \tau) \\ a_{min}(\Psi M) \\ b_{max}(\Psi M) \end{bmatrix}$$

where we define $a_{min}(\Psi M) = \min\{a(\tau)\colon \tau \in T(\delta)\}$, $b_{max}(\Psi M) = \max\{b(\tau)\colon \tau \in T(\delta)\}$, $T(\delta) = \left\{\tau\colon \Delta(\Psi M, \tau) \leq \chi^2_q(\delta)\right\}$ and $\Delta(\Psi M, \tau) = (\Psi M - \tau)' (\Psi \Omega \Psi')^{-1} (\Psi M - \tau)$. By the Continuous Mapping Theorem

$$\mathbb{P}\left[\widehat{a}_{min}(\widehat{\tau}) \leq \sqrt{n}\,(\widehat{\mu} - \mu_{true}) \leq \widehat{b}_{max}(\widehat{\tau})\right] \rightarrow \mathbb{P}\left[a_{min}(\Psi M) \leq \Lambda(\tau) \leq b_{max}(\Psi M)\right]$$

so it suffices to show that

$$\mathbb{P}\left[a_{min}(\Psi M) \leq \Lambda(\tau) \leq b_{max}(\Psi M)\right] \geq 1 - (\alpha + \delta).$$

Define the event $A = \left\{\Delta(\Psi M, \tau) \leq \chi^2_q(\delta)\right\}$, so that $\mathbb{P}(A) = 1 - \delta$. Then,

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \\ &= \mathbb{P}\left[\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \cap A\right] + \mathbb{P}\left[\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \cap A^c\right] \end{aligned}$$

By the definitions of $a_{min}(\Psi M)$, $b_{max}(\Psi M)$ and A,

$$\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \cap A \subseteq \left\{a_0(\Psi M) \leq \Lambda(\tau) \leq b_0(\Psi M)\right\}$$

hence

$$\mathbb{P}\left[\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \cap A\right] \leq \mathbb{P}\left\{a_{min}(\Psi M) \leq \Lambda(\tau) \leq b_{max}(\Psi M)\right\}$$

Further, since

$$\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \cap A^c \subseteq A^c$$

we have

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

Combining:

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left[\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \cap A\right] + \mathbb{P}\left[\left\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\right\} \cap A^c\right] \\ &\leq \mathbb{P}\left\{a_{min}(\Psi M) \leq \Lambda(\tau) \leq b_{max}(\Psi M)\right\} + \delta \end{aligned}$$

Therefore

$$\mathbb{P}\left\{a_{min}(\Psi M) \leq \Lambda(\tau) \leq b_{max}(\Psi M)\right\} \geq 1 - (\alpha + \delta).$$

□

Appendix B. Primitive Conditions for Assumption 2.2

In this section I derive primitive conditions for Assumptions 2.2 (iv), (v), (vii) and (viii) when the triangular array $\{Z_{ni}\}$ is independent and identically distributed within each row.

Lemma Appendix B.1 (Continuity and Convergence of Moments)**.** *Suppose that*

(a) f is almost-surely continuous

(b) There exists a random variable $Y(\theta)$ such that $|f^{(j)}(Z_{ni}, \theta)| \leq Y(\theta)$ for all j, n with $\mathbb{E}[\sup_{\theta \in \Theta} Y(\theta)^2] < \infty$

Then,

(i) $\lim_{n\to\infty} \mathbb{E}[f(Z_{ni}, \theta)] = \mathbb{E}[f(Z, \theta)]$,

(ii) $\lim_{n\to\infty} Var[f(Z_{ni}, \theta_0)] = Var[f(Z, \theta)]$,

(iii) and $\mathbb{E}[f(Z, \theta)]$ is continuous

for all $\theta \in \Theta$.

Proof. For (i), simply apply the Lebesgue Dominated Convergence Theorem element-wise to interchange limit and expectation. For (ii), express the typical element of $Var[f(Z_{ni}, \theta)]$ as

$$\{Var[f(Z_{ni}, \theta)]\}_{j,k} = \mathbb{E}\left[f^{(j)}(Z_{ni}, \theta) f^{(k)}(Z_{ni}, \theta)\right] - \mathbb{E}\left[f^{(j)}(Z_{ni}, \theta)\right] \mathbb{E}\left[f^{(k)}(Z_{ni}, \theta)\right]$$

Convergence of the second term to $\mathbb{E}\left[f^{(j)}(Z, \theta)\right] \mathbb{E}\left[f^{(k)}(Z, \theta)\right]$ follows from (a) and continuity. For the first term, write

$$\mathbb{E}\left[f^{(j)}(Z_{ni}, \theta) f^{(k)}(Z_{ni}, \theta)\right] \leq \mathbb{E}\left[f^{(j)}(Z_{ni}, \theta)^2\right] \mathbb{E}\left[f^{(k)}(Z_{ni}, \theta)^2\right]$$

by Cauchy-Schwartz and again apply Lebesgue Dominated Convergence.

For (iii), since Z is the almost-sure limit of $\{Z_{ni}\}$ and f is almost-surely continuous, $f(Z, \theta)$ is the almost-sure limit of $f(Z_{ni}, \theta)$. Thus, $Y(\theta)$ dominates the components of $f(Z, \theta)$ so we may again apply Lebesgue Dominated Convergence to find

$$\lim_{\theta \to \theta^*} \mathbb{E}[f(Z, \theta)] = \mathbb{E}\left[\lim_{\theta \to \theta^*} f(Z, \theta)\right] = \mathbb{E}[f(Z, \theta^*)]$$

establishing continuity. □

Lemma Appendix B.2 (Uniform WLLN)**.** *Suppose that*

(a) The triangular array $\{Z_{ni}\}$ is iid within each row

(b) There exists a random variable $Y(\theta)$ such that $|f^{(j)}(Z_{ni}, \theta)| \leq Y(\theta)$ for all n and $j \in \{1, \ldots, p+q\}$, with $\mathbb{E}[\sup_{\theta \in \Theta} Y(\theta)^2] < \infty$

(c) f is almost-surely differentiable on Θ^*, an open convex set containing Θ

(d) $\mathbb{E}\left[\sup_{\theta^* \in \Theta^*} \|\nabla_\theta f^{(j)}(Z_{ni}, \theta^*)\|\right] = O(1)$*, for all* $j \in \{1, 2, \ldots, p+q\}$

Then, $\sup_{\theta \in \Theta} \|f_n(\theta) - \mathbb{E}[f(Z, \theta)]\| \to_p 0$

Proof. It suffices to establish that

$$\sup_{\theta \in \Theta} \left| f_n^{(j)}(\theta) - \mathbb{E}[f^{(j)}(Z, \theta)] \right| \to_p 0$$

for each component $j \in \{1, 2, \ldots, p+q\}$. By Davidson (1994) Theorem 21.9, this is equivalent to pointwise convergence in probability and stochastic equicontinuity of $\{f_n\}$.

To show pointwise convergence, first combine Markov's Inequality and the fact that $\{Z_{ni}\}$ is iid in each row, yielding

$$\mathbb{P}\left(\left| f_n^{(j)}(\theta) - \mathbb{E}[f^{(j)}(Z, \theta)] \right| > \epsilon \right) \leq \frac{\mathbb{E}\left\{ \left| f^{(j)}(Z_{ni}, \theta) - \mathbb{E}[f^{(j)}(Z, \theta)] \right|^2 \right\}}{n\epsilon^2}$$

Now define

$$\begin{aligned} A &= \left| f^{(j)}(Z_{ni}, \theta) - \mathbb{E}[f^{(j)}(Z_{ni}, \theta)] \right| \\ B &= \left| \mathbb{E}[f^{(j)}(Z_{ni}, \theta)] - \mathbb{E}[f^{(j)}(Z, \theta)] \right| \end{aligned}$$

By the triangle inequality it suffices to show that

$$\lim_{n \to \infty} n^{-1}\mathbb{E}[A^2] = \lim_{n \to \infty} n^{-1}\mathbb{E}[AB] = \lim_{n \to \infty} n^{-1}\mathbb{E}[B^2] = 0$$

First

$$n^{-1}\mathbb{E}[A^2] = n^{-1}Var[f_n^{(j)}(Z_{ni}, \theta)] \leq n^{-1}\mathbb{E}[f_n^{(j)}(Z_{ni}, \theta)^2] \leq n^{-1}\mathbb{E}\left[\sup_{\theta \in \Theta} Y(\theta)^2 \right] \to 0$$

next

$$\begin{aligned} n^{-1}\mathbb{E}[AB] &= n^{-1}\mathbb{E}\left\{ \left| f^{(j)}(Z_{ni}, \theta) - \mathbb{E}[f^{(j)}(Z_{ni}, \theta)] \right| \right\} \left| \mathbb{E}[f^{(j)}(Z_{ni}, \theta)] - \mathbb{E}[f^{(j)}(Z, \theta)] \right| \\ &\leq 2n^{-1}\mathbb{E}\left[\left| f^{(j)}(Z_{ni}, \theta) \right| \right] \left| \mathbb{E}[f^{(j)}(Z_{ni}, \theta)] - \mathbb{E}[f^{(j)}(Z, \theta)] \right| \\ &\leq 2n^{-1}\mathbb{E}\left[\sup_{\theta \in \Theta} Y(\theta)^2 \right] \left| \mathbb{E}[f^{(j)}(Z_{ni}, \theta)] - \mathbb{E}[f^{(j)}(Z, \theta)] \right| \end{aligned}$$

and finally

$$n^{-1}\mathbb{E}[B^2] = n^{-1} \left| \mathbb{E}[f^{(j)}(Z_{ni}, \theta)] - \mathbb{E}[f^{(j)}(Z, \theta)] \right|^2 \to 0$$

where we have used the fact that $\lim_{n \to \infty} \mathbb{E}[f(Z_{ni}, \theta)] = \mathbb{E}[f(Z, \theta)]$ as implied by (a) via Lemma Appendix B.1.

To establish stochastic equicontinuity, we appeal to Davidson (1994) Theorem 21.11, under which it suffices to find a stochastic array $\{B_{ni}\}$ such that $\sum_{i=1}^n \mathbb{E}[B_{ni}] = O(1)$ and

$$\frac{1}{n} \left| f^{(j)}(Z_{ni}, \theta_1) - f^{(j)}(Z_{ni}, \theta_2) \right| \leq B_{ni} \|\theta_1 - \theta_2\|$$

almost surely for all $\theta_1, \theta_2 \in \Theta$. Now, by the mean-value theorem, for any $\theta_1, \theta_2 \in \Theta^*$ we can find a $\bar{\theta} \in \Theta$ such that

$$f^{(j)}(Z_{ni}, \theta_1) - f^{(j)}(Z_{ni}, \theta_2) = \nabla_\theta f(Z_{ni}, \bar{\theta})'(\theta_1 - \theta_2)$$

almost surely. Taking the absolute value of both sides and applying the Cauchy-Schwarz inequality,

$$\left|f^{(j)}(Z_{ni}, \theta_1) - f^{(j)}(Z_{ni}, \theta_2)\right| \leq \sup_{\theta^* \in \Theta^*} \|\nabla_\theta f^{(j)}(Z_{ni}, \theta^*)\| \cdot \|\theta_1 - \theta_2\|$$

almost surely for all $\theta_1, \theta_2 \in \Theta^*$, where B_{ni} does not depend on θ. Setting,

$$\{B_{ni}\} = \left\{ n^{-1} \sup_{\theta^* \in \Theta^*} \|\nabla_\theta f^{(j)}(Z_{ni}, \theta^*)\| \right\}$$

we have

$$\sum_{i=1}^{n} B_{ni} = \frac{1}{n} \sum_{i=1}^{n} \sup_{\theta^* \in \Theta^*} \|\nabla_\theta f^{(j)}(Z_{ni}, \theta^*)\| = \sup_{\theta^* \in \Theta^*} \|\nabla_\theta f^{(j)}(Z_{ni}, \theta^*)\|$$

□

Corollary Appendix B.1 (Uniform WLLN for Derivative Matrix)**.** *Suppose that*

(a) The triangular array $\{Z_{ni}\}$ *is iid within each row*

(b) There exists a random variable $Y(\theta)$ *such that* $\left|\nabla_{\theta_k} f^{(j)}(Z_{ni}, \theta)\right| \leq Y(\theta)$ *for all* j, k, n *and* $\mathbb{E}[\sup_{\theta \in \Theta} Y(\theta)^2] < \infty$

(c) f *is twice differentiable almost-surely on* Θ^**, an open convex set containing* Θ

(d) $\mathbb{E}\left[\sup_{\theta^* \in \Theta^*} \|\nabla^2_{\theta, \theta_k} f^{(j)}(Z_{ni}, \theta^*)\|\right] = O(1)$*, for all* j, k

Then,

$$\sup_{\theta \in \Theta} \|\nabla_\theta f_n(\theta) - F(\theta)\| \to_p 0$$

Lemma Appendix B.3. *Suppose that*

(a) The triangular array $\{Z_{ni}\}$ *is iid within each row*

(b) $\lim_{n \to \infty} Var[f(Z_{ni}, \theta_0)] = \Omega$

(c) There is a random variable Y *such that* $\|f(Z_{ni}, \theta_0)\| \leq Y$ *for all* n *and* $\mathbb{E}[Y^{2+\delta}] < \infty$ *for some* $\delta > 0$*.*

Then,

$$\sqrt{n} f_n(\theta_0) \to_d \mathcal{N}_{p+q}\left(\begin{bmatrix} 0 \\ \tau \end{bmatrix}, \Omega \right)$$

Proof. Define $A_n(\epsilon) = \{\|f(Z_{ni}, \theta_0)\| > \epsilon\sqrt{n}\}$. By Hölder's Inequality,

$$\begin{aligned}\mathbb{E}\left[\|f(Z_{ni},\theta_0)\|^2 \mathbf{1}\left\{A_n(\epsilon)\right\}\right] &\leq \left\{\mathbb{E}\left[\|f(Z_{ni},\theta_0)\|^{2+\delta}\right]\right\}^{2/(2+\delta)} \left\{\mathbb{E}\left[\mathbf{1}\left\{A_n(\epsilon)\right\}^{(2+\delta)/\delta}\right]\right\}^{\delta/(2+\delta)} \\ &= \left\{\mathbb{E}\left[\|f(Z_{ni},\theta_0)\|^{2+\delta}\right]\right\}^{2/(2+\delta)} \left[\mathbb{P}\left\{A_n(\epsilon)\right\}\right]^{\delta/(2+\delta)}\end{aligned}$$

and by Markov's Inequality

$$\mathbb{P}\left\{A_n(\epsilon)\right\} = \mathbb{P}\left\{\|f(Z_{ni},\theta_0)\| > \epsilon\sqrt{n}\right\} \leq \frac{\mathbb{E}\left[\|f(Z_{ni},\theta_0)\|\right]}{\epsilon\sqrt{n}}$$

Since $\|f(Z_{ni},\theta_0)\| \leq Y$,

$$\mathbb{E}[\|f(Z_{ni},\theta_0)\|^{2+\delta}] \leq \mathbb{E}[Y^{2+\delta}] < \infty$$

and

$$\mathbb{E}[\|f(Z_{ni},\theta_0)\|] \leq \mathbb{E}[Y] < \infty$$

Combining,

$$\mathbb{E}\left[\|f(Z_{ni},\theta_0)\|^2 \mathbf{1}\left\{A_n(\epsilon)\right\}\right] \leq \left\{\mathbb{E}[Y^{2+\delta}]\right\}^{2/(2+\delta)} \left\{\frac{\mathbb{E}\left[Y\right]}{\epsilon\sqrt{n}}\right\}^{\delta/(2+\delta)}$$

so that

$$\lim_{n\to\infty} \mathbb{E}\left[\|f(Z_{ni},\theta_0)\|^2 \mathbf{1}\left\{\|f(Z_{ni},\theta_0)\| > \epsilon\sqrt{n}\right\}\right] = 0$$

Thus, by the Lindeberg-Feller Central Limit Theorem,

$$\sqrt{n} f_n(\theta_0) - \sqrt{n}\,\mathbb{E}[f(Z_{ni},\theta_0)] \to_d \mathcal{N}_{p+q}(0,\Omega)$$

Now, by Assumption 2.1

$$\sqrt{n}\mathbb{E}[f(Z_{ni},\theta_0)] = \begin{bmatrix} 0 \\ \tau \end{bmatrix}$$

so that

$$\sqrt{n} f_n(\theta_0) \to_d \mathcal{N}_{p+q}\left(\begin{bmatrix} 0 \\ \tau \end{bmatrix}, \Omega\right)$$

as asserted. □

Theorem Appendix B.1 (Primitive Conditions for Assumption 2.2)**.** *Suppose that*

(a) θ_0 *lies in the interior of* Θ*, a compact set*

(b) $\widetilde{W} \to_p W$*, a positive semi-definite matrix*

(c) $W\mathbb{E}[f(Z,\theta)] = 0$ *and* $W_{gg}\mathbb{E}[g(Z,\theta)] = 0$ *if and only if* $\theta = \theta_0$

(d) The triangular array $\{Z_{ni}\}$ *is iid in each row*

(e) f *is almost surely twice differentiable in an open convex set* Θ^* *containing* Θ

(f) There is a random variable $Y(\theta)$ *that dominates* $|f^{(j)}(Z_{ni},\theta)|$, $\left|\nabla_{\theta_k} f^{(j)}(Z_{ni},\theta)\right|$ *and* $\left|\nabla^2_{\theta_j,\theta_k} f^{(j)}(Z_{ni},\theta)\right|$ *for all* j,k,n *where* $\mathbb{E}\left[\sup_{\theta\in\Theta} Y(\theta)^{2+\delta}\right] < \infty$

(g) $F'WF$ *and* $G'W_{gg}G$ *are invertible*

Then Assumption 2.2 is satisfied.

Proof. Conditions (d)–(f) imply Assumptions 2.2 (d)–(h) by the preceding lemmas. The remaining conditions are simply Assumptions 2.2 (a)–(c) and (i). □

Appendix C. Supplementary Tables and Figures

Table C.22: Difference in RMSE between the estimator including w (full) and the estimator excluding it (valid) over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument. Negative values indicate that including w gives a smaller RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 50.

	$\rho = Cov(w,u)$								
$N = 50$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.14	-0.14	-0.09	-0.05	0.00	0.03	0.12	0.15	0.20
0.1	-0.15	-0.15	-0.13	-0.07	-0.01	0.03	0.10	0.14	0.22
0.2	-0.25	-0.23	-0.19	-0.17	-0.04	0.03	0.00	0.10	0.24
0.3	-0.37	-0.31	-0.36	-0.21	-0.16	-0.11	0.01	-0.01	0.14
0.4	-0.44	-0.39	-0.34	-0.27	-0.25	-0.14	-0.15	0.00	0.07
0.5	-0.44	-0.41	-0.40	-0.37	-0.29	-0.22	-0.16	-0.09	0.00
0.6	-0.45	-0.45	-0.42	-0.38	-0.34	-0.32	-0.30	-0.15	-0.08
0.7	-0.47	-0.46	-0.43	-0.47	-0.32	-0.30	-0.25	-0.19	-0.16
0.8	-0.47	-0.47	-0.46	-0.45	-0.36	-0.35	-0.26	-0.20	-0.19
0.9	-0.52	-0.46	-0.54	-0.40	-0.36	-0.34	-0.47	-0.22	-0.19
1.0	-0.49	-0.45	-0.41	-0.44	-0.39	-0.34	-0.31	-0.26	-0.20
1.1	-0.46	-0.48	-0.44	-0.42	-0.37	-0.41	-0.34	-0.24	-0.21
1.2	-0.50	-0.44	-0.43	-0.39	-0.38	-0.34	-0.29	-0.29	-0.24
1.3	-0.43	-0.42	-0.42	-0.40	-0.43	-0.33	-0.32	-0.25	-0.22

$\gamma = Cov(w, x)$

Table C.23: Difference in RMSE between the estimator including w (full) and the estimator excluding it (valid) over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument. Negative values indicate that including w gives a smaller RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 100.

	$\rho = Cov(w,u)$								
$N = 100$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.10	-0.08	-0.05	-0.01	0.06	0.16	0.22	0.25	0.39
0.1	-0.14	-0.13	-0.08	0.00	0.10	0.17	0.27	0.36	0.49
0.2	-0.41	-0.25	-0.14	-0.09	0.02	0.12	0.26	0.29	0.45
0.3	-0.30	-0.34	-0.24	-0.15	-0.08	0.04	0.12	0.25	0.36
0.4	-0.50	-0.36	-0.29	-0.23	-0.23	-0.08	0.03	0.07	0.23
0.5	-0.43	-0.39	-0.35	-0.31	-0.66	-0.12	-0.05	0.03	0.11
0.6	-0.44	-0.39	-0.35	-0.37	-0.24	-0.21	-0.16	-0.04	-0.05
0.7	-0.46	-0.42	-0.38	-0.35	-0.31	-0.28	-0.17	-0.11	-0.03
0.8	-0.45	-0.46	-0.45	-0.38	-0.31	-0.25	-0.19	-0.16	-0.06
0.9	-0.43	-0.43	-0.45	-0.35	-0.33	-0.24	-0.26	-0.16	-0.12
1.0	-0.44	-0.42	-0.43	-0.38	-0.34	-0.27	-0.24	-0.18	-0.16
1.1	-0.43	-0.42	-0.38	-0.39	-0.32	-0.30	-0.23	-0.21	-0.21
1.2	-0.43	-0.43	-0.40	-0.36	-0.33	-0.34	-0.27	-0.21	-0.23
1.3	-0.41	-0.42	-0.47	-0.35	-0.33	-0.30	-0.29	-0.22	-0.19

$\gamma = Cov(w, x)$

Table C.24: Difference in RMSE between the estimator selected by the FMSC and that selected by the GMM-BIC over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w, u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	-0.01	-0.02	-0.04	-0.05	-0.05	-0.07	-0.07	-0.08
0.1	0.00	-0.01	-0.05	-0.05	-0.04	-0.04	-0.04	-0.05	-0.08
0.2	0.03	0.00	-0.04	-0.09	-0.10	-0.09	-0.09	-0.08	-0.09
0.3	0.04	0.02	-0.02	-0.07	-0.10	-0.13	-0.13	-0.12	-0.12
0.4	0.06	0.04	0.00	-0.04	-0.08	-0.12	-0.14	-0.14	-0.15
0.5	0.06	0.04	0.01	-0.02	-0.05	-0.09	-0.12	-0.15	-0.17
0.6	0.05	0.04	0.03	0.00	-0.03	-0.06	-0.10	-0.13	-0.16
0.7	0.06	0.05	0.02	-0.01	-0.03	-0.05	-0.07	-0.10	-0.13
0.8	0.04	0.04	0.02	0.02	-0.01	-0.03	-0.05	-0.08	-0.11
0.9	0.06	0.04	0.03	0.01	-0.01	-0.02	-0.04	-0.06	-0.09
1.0	0.03	0.02	0.01	0.02	0.00	-0.02	-0.04	-0.05	-0.06
1.1	0.06	0.05	0.03	0.01	0.00	-0.01	-0.02	-0.03	-0.05
1.2	0.04	0.04	0.03	0.01	0.01	0.00	-0.03	-0.03	-0.04
1.3	0.04	0.03	0.03	0.00	0.00	0.00	-0.01	-0.03	-0.04

$\gamma = Cov(w, x)$

Table C.25: Difference in RMSE between the estimator selected by the FMSC and that selected by the GMM-HQ over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w, u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	-0.01	-0.02	-0.03	-0.03	-0.04	-0.04	-0.05	-0.05
0.1	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.05
0.2	0.01	-0.01	-0.02	-0.04	-0.05	-0.03	-0.04	-0.04	-0.04
0.3	0.01	0.00	-0.01	-0.03	-0.05	-0.06	-0.05	-0.05	-0.05
0.4	0.02	0.01	0.00	-0.02	-0.04	-0.06	-0.06	-0.06	-0.06
0.5	0.01	0.00	-0.01	-0.02	-0.02	-0.04	-0.05	-0.07	-0.08
0.6	0.01	0.00	0.00	-0.01	-0.02	-0.04	-0.05	-0.07	-0.07
0.7	0.01	0.00	-0.01	-0.02	-0.03	-0.03	-0.04	-0.05	-0.06
0.8	0.00	-0.01	-0.02	-0.01	-0.02	-0.02	-0.03	-0.04	-0.06
0.9	0.00	-0.01	-0.01	-0.02	-0.02	-0.02	-0.03	-0.04	-0.06
1.0	-0.03	-0.03	-0.03	-0.01	-0.02	-0.03	-0.03	-0.04	-0.04
1.1	-0.01	-0.02	-0.04	-0.04	-0.02	-0.03	-0.03	-0.03	-0.04
1.2	-0.03	-0.01	-0.02	-0.03	-0.02	-0.02	-0.04	-0.04	-0.04
1.3	-0.06	-0.02	-0.03	-0.04	-0.04	-0.03	-0.03	-0.03	-0.04

$\gamma = Cov(w, x)$

Table C.26: Difference in RMSE between the estimator selected by the FMSC and that selected by the GMM-AIC over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w, u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.03
0.1	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.03
0.2	-0.02	-0.02	0.00	0.00	-0.02	-0.01	-0.02	-0.02	-0.02
0.3	-0.02	-0.02	0.00	0.00	0.00	-0.02	-0.02	-0.02	-0.02
0.4	-0.03	-0.02	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.02
0.5	-0.03	-0.03	-0.01	0.00	0.01	0.01	0.00	-0.02	-0.02
0.6	-0.03	-0.03	-0.02	-0.01	0.00	0.00	0.00	-0.01	-0.02
0.7	-0.04	-0.04	-0.04	-0.03	-0.02	0.00	0.00	0.00	-0.01
0.8	-0.04	-0.05	-0.04	-0.02	-0.02	-0.01	0.00	0.00	-0.02
0.9	-0.04	-0.05	-0.04	-0.04	-0.03	-0.02	-0.01	-0.01	-0.02
1.0	-0.07	-0.07	-0.07	-0.04	-0.03	-0.03	-0.02	-0.02	-0.01
1.1	-0.06	-0.06	-0.07	-0.06	-0.04	-0.04	-0.03	-0.02	-0.02
1.2	-0.08	-0.06	-0.07	-0.06	-0.05	-0.04	-0.04	-0.03	-0.03
1.3	-0.11	-0.07	-0.08	-0.07	-0.06	-0.04	-0.04	-0.03	-0.03

$\gamma = Cov(w, x)$

Table C.27: Difference in RMSE between the estimator selected by the FMSC and that selected by the CC-MSC-BIC over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w, u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.1	-0.02	-0.02	-0.03	-0.03	-0.03	-0.02	-0.02	-0.03	-0.04
0.2	0.03	0.00	-0.04	-0.08	-0.11	-0.09	-0.09	-0.08	-0.09
0.3	0.04	0.02	-0.02	-0.07	-0.10	-0.13	-0.13	-0.12	-0.12
0.4	0.06	0.04	0.00	-0.04	-0.08	-0.12	-0.14	-0.14	-0.15
0.5	0.06	0.04	0.01	-0.02	-0.05	-0.09	-0.12	-0.15	-0.17
0.6	0.05	0.04	0.03	0.00	-0.03	-0.06	-0.10	-0.13	-0.16
0.7	0.06	0.05	0.02	-0.01	-0.03	-0.05	-0.07	-0.10	-0.13
0.8	0.04	0.04	0.02	0.02	-0.01	-0.03	-0.05	-0.08	-0.11
0.9	0.06	0.04	0.03	0.01	-0.01	-0.02	-0.04	-0.06	-0.09
1.0	0.03	0.02	0.01	0.02	0.00	-0.02	-0.04	-0.05	-0.06
1.1	0.06	0.05	0.03	0.01	0.00	-0.01	-0.02	-0.03	-0.05
1.2	0.04	0.04	0.03	0.01	0.01	0.00	-0.03	-0.03	-0.04
1.3	0.04	0.03	0.03	0.00	0.00	0.00	-0.01	-0.03	-0.04

$\gamma = Cov(w, x)$

Table C.28: Difference in RMSE between the estimator selected by the FMSC and that selected by the CC-MSC-HQ over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.00
0.1	-0.01	-0.02	-0.02	-0.02	-0.01	-0.01	-0.02	-0.02	-0.04
0.2	0.01	-0.01	-0.02	-0.03	-0.05	-0.03	-0.04	-0.03	-0.04
0.3	0.01	0.00	-0.01	-0.03	-0.05	-0.06	-0.05	-0.05	-0.05
0.4	0.02	0.01	0.00	-0.02	-0.04	-0.06	-0.06	-0.06	-0.06
0.5	0.01	0.00	-0.01	-0.02	-0.02	-0.04	-0.05	-0.07	-0.08
0.6	0.01	0.00	0.00	-0.01	-0.02	-0.04	-0.05	-0.07	-0.07
0.7	0.01	0.00	-0.01	-0.02	-0.03	-0.03	-0.04	-0.05	-0.06
0.8	0.00	-0.01	-0.02	-0.01	-0.02	-0.02	-0.03	-0.04	-0.06
0.9	0.00	-0.01	-0.01	-0.02	-0.02	-0.02	-0.03	-0.04	-0.06
1.0	-0.03	-0.03	-0.03	-0.01	-0.02	-0.03	-0.03	-0.04	-0.04
1.1	-0.01	-0.02	-0.04	-0.04	-0.02	-0.03	-0.03	-0.03	-0.04
1.2	-0.03	-0.01	-0.02	-0.03	-0.02	-0.02	-0.04	-0.04	-0.04
1.3	-0.06	-0.02	-0.03	-0.04	-0.04	-0.03	-0.03	-0.03	-0.04

$\gamma = Cov(w, x)$

Table C.29: Difference in RMSE between the estimator selected by the FMSC and that selected by the CC-MSC-AIC over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with 10,000 replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.1	-0.02	-0.01	0.00	0.00	-0.01	-0.01	-0.01	-0.02	-0.03
0.2	-0.02	-0.02	0.00	0.00	-0.02	-0.01	-0.02	-0.02	-0.02
0.3	-0.02	-0.02	0.00	0.00	0.00	-0.02	-0.02	-0.02	-0.02
0.4	-0.03	-0.02	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.02
0.5	-0.03	-0.03	-0.01	0.00	0.01	0.01	0.00	-0.02	-0.02
0.6	-0.03	-0.03	-0.02	-0.01	0.00	0.00	0.00	-0.01	-0.02
0.7	-0.04	-0.04	-0.04	-0.03	-0.02	0.00	0.00	0.00	-0.01
0.8	-0.04	-0.05	-0.04	-0.02	-0.02	-0.01	0.00	0.00	-0.02
0.9	-0.04	-0.05	-0.04	-0.04	-0.03	-0.02	-0.01	-0.01	-0.02
1.0	-0.07	-0.07	-0.07	-0.04	-0.03	-0.03	-0.02	-0.02	-0.01
1.1	-0.06	-0.06	-0.07	-0.06	-0.04	-0.04	-0.03	-0.02	-0.02
1.2	-0.08	-0.06	-0.07	-0.06	-0.05	-0.04	-0.04	-0.03	-0.03
1.3	-0.11	-0.07	-0.08	-0.07	-0.06	-0.04	-0.04	-0.03	-0.03

$\gamma = Cov(w, x)$

Table C.30: Difference in RMSE between the estimator selected by the FMSC and that selected by a downward J-test at the 90% level over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	-0.01	-0.02	-0.03	-0.04	-0.04	-0.06	-0.05	-0.06
0.1	-0.01	-0.01	-0.02	-0.03	-0.02	-0.02	-0.02	-0.03	-0.05
0.2	0.02	0.00	-0.02	-0.05	-0.05	-0.04	-0.04	-0.04	-0.04
0.3	0.02	0.01	-0.01	-0.04	-0.06	-0.07	-0.07	-0.06	-0.06
0.4	0.03	0.02	0.00	-0.02	-0.05	-0.07	-0.08	-0.07	-0.08
0.5	0.03	0.02	0.00	-0.01	-0.03	-0.05	-0.07	-0.09	-0.10
0.6	0.02	0.02	0.01	0.00	-0.02	-0.04	-0.06	-0.08	-0.09
0.7	0.03	0.02	0.00	-0.01	-0.02	-0.03	-0.04	-0.06	-0.08
0.8	0.01	0.01	0.00	0.01	-0.01	-0.02	-0.03	-0.05	-0.07
0.9	0.03	0.01	0.01	0.00	-0.01	-0.02	-0.03	-0.04	-0.06
1.0	0.00	0.00	-0.01	0.01	0.00	-0.02	-0.03	-0.04	-0.04
1.1	0.02	0.02	0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.04
1.2	0.00	0.02	0.00	-0.01	-0.01	-0.01	-0.03	-0.02	-0.03
1.3	0.01	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.03

$\gamma = Cov(w, x)$

Table C.31: Difference in RMSE between the estimator selected by the FMSC and that selected by a downward J-test at the 95% level over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Negative values indicate that the FMSC gives a lower realized RMSE. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

	$\rho = Cov(w,u)$								
$N = 500$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.0	0.00	-0.01	-0.02	-0.04	-0.05	-0.05	-0.07	-0.07	-0.08
0.1	0.00	-0.01	-0.04	-0.04	-0.03	-0.03	-0.03	-0.04	-0.06
0.2	0.03	0.00	-0.03	-0.08	-0.09	-0.07	-0.07	-0.06	-0.07
0.3	0.04	0.02	-0.01	-0.06	-0.09	-0.12	-0.11	-0.11	-0.10
0.4	0.05	0.03	0.01	-0.03	-0.07	-0.11	-0.13	-0.13	-0.13
0.5	0.05	0.04	0.02	-0.01	-0.04	-0.08	-0.11	-0.14	-0.15
0.6	0.05	0.04	0.03	0.01	-0.02	-0.06	-0.09	-0.12	-0.14
0.7	0.05	0.04	0.02	0.01	-0.02	-0.03	-0.06	-0.09	-0.12
0.8	0.04	0.04	0.03	0.02	0.00	-0.02	-0.05	-0.07	-0.10
0.9	0.06	0.03	0.03	0.02	0.00	-0.02	-0.03	-0.05	-0.08
1.0	0.03	0.04	0.03	0.02	0.01	0.00	-0.03	-0.04	-0.06
1.1	0.06	0.05	0.03	0.01	0.00	-0.01	-0.02	-0.03	-0.04
1.2	0.04	0.04	0.03	0.01	0.02	0.00	-0.02	-0.02	-0.04
1.3	0.04	0.03	0.03	0.01	0.01	0.01	-0.01	-0.02	-0.03

$\gamma = Cov(w, x)$

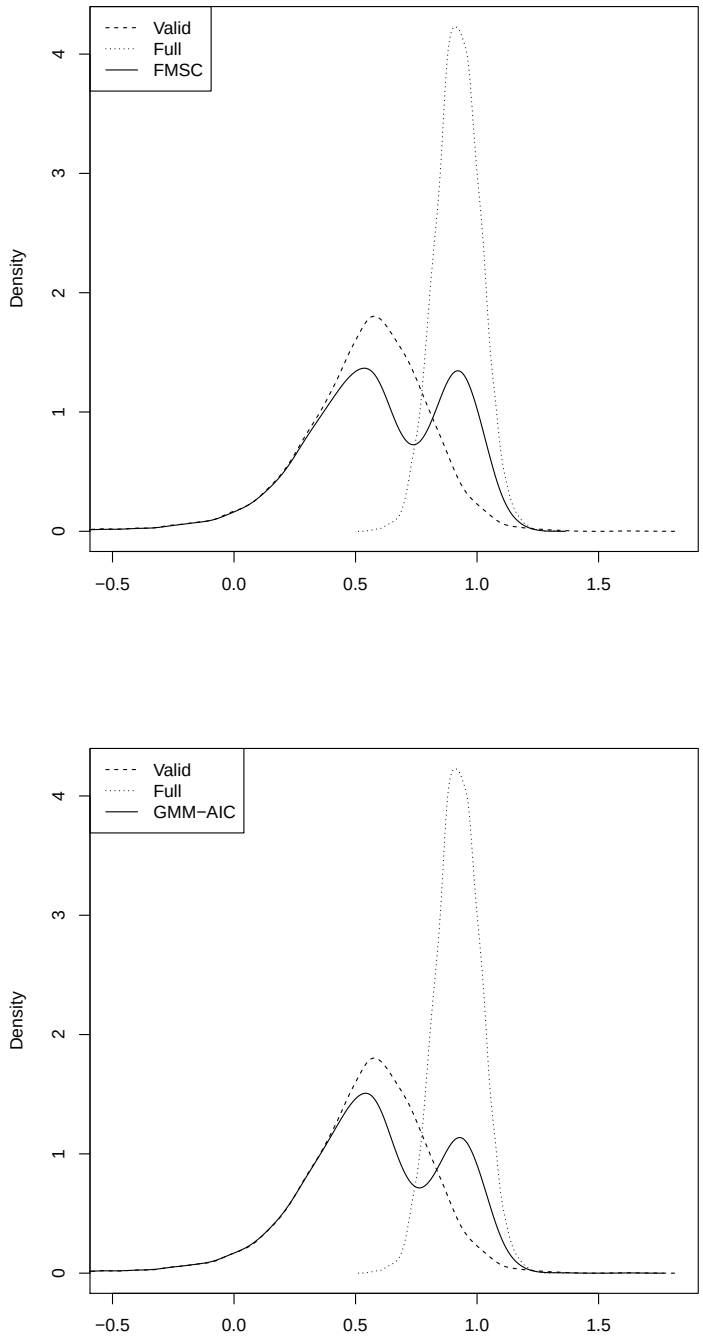

Figure C.2: Post-selection distributions for the estimated effect of x on y in Equation 3.17 with $\gamma = 0.4$, $\rho = 0.2$, $N = 500$. The distribution post-FMSC selection appears in the top panel, while the distribution post-GMM-AIC selection appears in the bottom panel. The distribution of the full estimator is given in dotted lines while that of the valid estimator is given in dashed lines in each panel. All distributions are calculated by kernel density estimation based on 10,000 simulation replications generated from Equations 3.17–3.19.

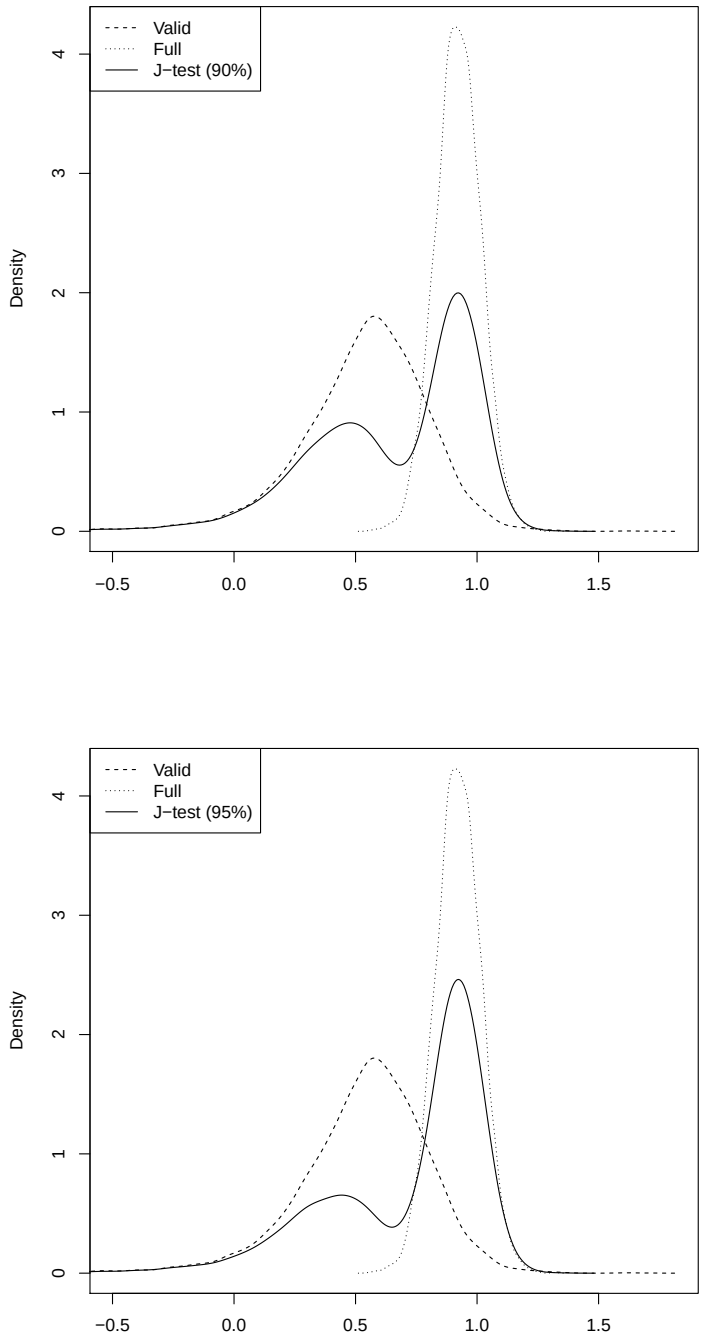

Figure C.3: Post-selection distributions for the estimated effect of x on y in Equation 3.17 with $\gamma = 0.4$, $\rho = 0.2$, $N = 500$. The distribution after a downward J-test at the 90% level appears in the top panel, while the distribution after a downward J-test at the 95% level appears in the bottom panel. The distribution of the full estimator is given in dotted lines while that of the valid estimator is given in dashed lines in each panel. All distributions are calculated by kernel density estimation based on 10,000 simulation replications generated from Equations 3.17–3.19.

Table C.32: Coverage probabilities after selection using a downward J-test at the 90% level of a traditional 95% asymptotic confidence interval for the effect of x on y in Equation 3.17, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

		$\rho = Cov(w,u)$								
$N = 500$		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w,x)$	0.0	0.92	0.92	0.93	0.93	0.93	0.92	0.93	0.92	0.93
	0.1	0.92	0.86	0.83	0.88	0.92	0.93	0.93	0.92	0.93
	0.2	0.92	0.79	0.65	0.70	0.83	0.90	0.91	0.92	0.92
	0.3	0.92	0.77	0.54	0.51	0.65	0.80	0.87	0.89	0.91
	0.4	0.92	0.77	0.48	0.37	0.47	0.64	0.75	0.83	0.87
	0.5	0.92	0.77	0.45	0.30	0.36	0.49	0.62	0.73	0.79
	0.6	0.92	0.77	0.44	0.25	0.28	0.38	0.48	0.59	0.68
	0.7	0.92	0.77	0.42	0.22	0.24	0.31	0.40	0.49	0.58
	0.8	0.91	0.77	0.43	0.21	0.20	0.26	0.32	0.41	0.49
	0.9	0.93	0.76	0.42	0.20	0.18	0.23	0.28	0.34	0.41
	1.0	0.92	0.78	0.42	0.18	0.16	0.19	0.23	0.30	0.35
	1.1	0.92	0.78	0.42	0.19	0.15	0.18	0.22	0.25	0.31
	1.2	0.93	0.78	0.42	0.19	0.14	0.16	0.20	0.23	0.27
	1.3	0.93	0.78	0.43	0.17	0.13	0.15	0.17	0.21	0.24

Table C.33: Coverage probabilities after selection using a downward J-test at the 95% level of a traditional 95% asymptotic confidence interval for the effect of x on y in Equation 3.17, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

		$\rho = Cov(w,u)$								
$N = 500$		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w,x)$	0.0	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.93
	0.1	0.92	0.84	0.79	0.84	0.91	0.92	0.93	0.92	0.93
	0.2	0.92	0.78	0.59	0.60	0.76	0.87	0.90	0.91	0.91
	0.3	0.93	0.76	0.48	0.40	0.54	0.71	0.81	0.86	0.89
	0.4	0.93	0.76	0.43	0.27	0.35	0.51	0.66	0.75	0.82
	0.5	0.93	0.76	0.41	0.22	0.25	0.36	0.50	0.62	0.70
	0.6	0.93	0.76	0.40	0.18	0.18	0.26	0.36	0.46	0.56
	0.7	0.93	0.76	0.39	0.16	0.15	0.20	0.28	0.36	0.46
	0.8	0.92	0.76	0.39	0.15	0.12	0.16	0.22	0.29	0.36
	0.9	0.94	0.76	0.39	0.14	0.10	0.14	0.18	0.23	0.29
	1.0	0.93	0.77	0.39	0.13	0.09	0.11	0.15	0.20	0.24
	1.1	0.93	0.77	0.39	0.13	0.09	0.11	0.13	0.16	0.20
	1.2	0.93	0.77	0.39	0.13	0.08	0.09	0.12	0.14	0.17
	1.3	0.94	0.78	0.40	0.12	0.07	0.08	0.10	0.13	0.15

Table C.34: Coverage probabilities post-GMM-AIC moment selection of a traditional 95% asymptotic confidence interval for the effect of x on y in Equation 3.17, over a grid of values for the relevance, $Cov(w, x)$, and validity, $Cov(w, u)$, of the instrument w. Values are calculated by simulating from Equations 3.17–3.19 with $10,000$ replications and a sample size of 500.

		$\rho = Cov(w, u)$								
$N = 500$		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\gamma = Cov(w, x)$	0.0	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	0.1	0.92	0.88	0.89	0.92	0.93	0.93	0.93	0.93	0.93
	0.2	0.91	0.81	0.76	0.84	0.91	0.92	0.92	0.92	0.92
	0.3	0.91	0.78	0.64	0.70	0.82	0.89	0.91	0.92	0.92
	0.4	0.90	0.77	0.56	0.55	0.69	0.82	0.88	0.90	0.91
	0.5	0.90	0.76	0.52	0.45	0.57	0.71	0.80	0.86	0.89
	0.6	0.90	0.76	0.50	0.39	0.47	0.60	0.70	0.79	0.83
	0.7	0.90	0.76	0.47	0.34	0.41	0.52	0.62	0.70	0.77
	0.8	0.89	0.76	0.47	0.32	0.36	0.44	0.55	0.63	0.70
	0.9	0.91	0.76	0.46	0.30	0.31	0.39	0.48	0.56	0.63
	1.0	0.91	0.77	0.46	0.27	0.29	0.35	0.43	0.51	0.56
	1.1	0.91	0.78	0.46	0.27	0.27	0.33	0.40	0.45	0.51
	1.2	0.92	0.78	0.47	0.27	0.26	0.30	0.36	0.41	0.47
	1.3	0.92	0.79	0.48	0.26	0.26	0.29	0.33	0.38	0.44