

Lecture 6: Moment Selection for GMM

Francis J. DiTraglia

April 3, 2014

1 Review of Generalized Method of Moments

The best all-around reference for for GMM is Hall (2005). These notes draw on chapters 3–7 of his book and use essentially the same notation.

1.1 Key Assumptions

Let f be a q -vector of functions of an observable random r -vector v_t and a p -vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^p$ where Θ is compact. The GMM estimator is defined as follows:

$$\begin{aligned}\bar{g}_T(\theta) &= \frac{1}{T} \sum_{t=1}^T f(v_t, \theta) \\ \hat{\theta}_T &= \arg \min_{\theta \in \Theta} \bar{g}_T(\theta)' W_T \bar{g}_T(\theta)\end{aligned}$$

The basic assumptions required for GMM estimation are as follows.

Strict Stationarity The sequence $\{v_t: -\infty < t < \infty\}$ of random r -vectors is a strictly stationary process with sample space $\mathcal{V} \subseteq \mathbb{R}^r$. Importantly, this implies that the expectations of *any* functions of v_t are constant over time.

Population Moment Condition $E[f(v_t, \theta_0)] = 0$ for some $\theta_0 \in \text{interior}(\Theta)$.

Global Identification For any $\tilde{\theta} \in \Theta$ such that $\tilde{\theta} \neq \theta_0$, $E[f(v_t, \tilde{\theta})] \neq 0$.

Weighting Matrix The $(q \times q)$ weighting matrix W_T is positive semi-definite and converges in probability to a positive definite constant matrix W .

1.2 Regularity Conditions

Regularity Conditions for Moment Functions The q moment functions $f: \mathcal{V} \times \Theta \rightarrow \mathbb{R}^q$ satisfy the following conditions:

- (i) f is v_t -almost surely continuous on Θ
- (ii) $E[f(v_t, \theta)] < \infty$ exists and is continuous on Θ

Regularity Conditions for Derivative Matrix

- (i) The $q \times p$ matrix $\nabla_{\theta'} f(v_t, \theta)$ exists and is v_t -almost continuous on Θ
- (ii) $E[\nabla_{\theta} f(v_t, \theta_0)] < \infty$ exists and is continuous in a neighborhood N_ϵ of θ_0
- (iii) $\sup_{\theta \in N_\epsilon} \left\| T^{-1} \sum_{t=1}^T \nabla_{\theta} f(v_t, \theta) - E[\nabla_{\theta} f(v_t, \theta)] \right\| \xrightarrow{p} 0$

Regularity Conditions for Variance of Sample Moment Conditions

- (i) $E[f(v_t, \theta_0)f(v_t, \theta_0)']$ exists and is finite.
- (ii) $\lim_{T \rightarrow \infty} \text{Var} \left[\sqrt{T} \bar{g}_T(\theta_0) \right] = S$ exists and is a finite, positive definite matrix.

1.3 Asymptotics Under Correct Specification

Under the set of assumptions given above, we obtain the following:

Consistency of GMM Estimator $\hat{\theta}_T \xrightarrow{p} \theta_0$

Asymptotic Normality of GMM Estimator $\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(0, MSM')$

$$\begin{aligned} M &= (G_0' W G_0)^{-1} G_0' W \\ G_0 &= E[\nabla_{\theta'} f(v_t, \theta_0)] \end{aligned}$$

1.4 The J-test Statistic

The J -test statistic is given by

$$J_T = T \bar{g}_T(\hat{\theta}_T)' \hat{S}^{-1} \bar{g}_T(\hat{\theta}_T)$$

where \hat{S} is a consistent estimator of S , the long-run variance matrix of the GMM sample moment conditions. We will need to consider the asymptotic behavior of this quantity in two settings: when the population moment condition is satisfied, and when it is violated.

Correct Specification Earlier in this document we reviewed the basic asymptotic results for GMM estimation under standard regularity conditions *assuming the population moment condition is correct*. Our main findings were that, regardless of weighting matrix, GMM is consistent and asymptotically normal. The particular choice of W_T *only* affects the asymptotic variance of the estimator. To study the behavior of the J -test in this setting, we need to examine an asymptotic expansion for the estimated sample moment. Using Taylor Expansion arguments, we can show that

$$W_T^{1/2} \sqrt{T} \bar{g}_T(\hat{\theta}_T) = [I_q - P(\theta_0)] W_T^{1/2} \sqrt{T} \bar{g}_T(\theta_0) + o_p(1)$$

where

$$\begin{aligned} P(\theta_0) &= F(\theta_0) [F(\theta_0)' F(\theta_0)]^{-1} F(\theta_0)' \\ F(\theta_0) &= W_T^{1/2} E[\nabla_{\theta} f(v_t, \theta_0)] \end{aligned}$$

The matrix $P(\theta_0)$ is called the *identifying restrictions* and corresponds to the particular projection of $W^{1/2}E[f(v_t, \theta)]$ actually used in GMM estimation. Its orthogonal complement, $N = I_q - P(\theta_0)$, is called the *overidentifying restrictions*. The expansion just stated shows that the asymptotic behavior of the estimated sample moment is *entirely governed by the overidentifying restrictions*. Via a CLT for $\sqrt{T}\bar{g}_T(\theta_0)$, it follows that

$$W_T^{1/2}\sqrt{T}\bar{g}_T(\hat{\theta}_T) \xrightarrow{d} \mathcal{N}(0, NW^{1/2}SW^{1/2}N')$$

Note that the $N = I_q - P(\theta_0)$ has rank $q - p$ since it is the orthogonal complement of the rank p projection matrix $P(\theta_0)$. Hence, in the limit we obtain a *singular normal distribution*, that is a q -dimensional random vector that concentrates on a $(q - p)$ -dimensional subspace of \mathbb{R}^q . Substituting the efficient weighting matrix \hat{S}^{-1} we find that $J_T \xrightarrow{d} \chi_{q-p}^2$ by the Continuous Mapping Theorem, *assuming that the population moment condition is correct*.

Incorrect Specification When the GMM population moment condition $E[f(v_t, \theta)] = 0$ is *false* for all $\theta \in \Theta$, the situation is completely different. In this case the probability limit of $\hat{\theta}_T$ in general *will* depend on the choice of weighting matrix and the rate of convergence depends on the rate at which W_T converges to W . Unsurprisingly, this leads to very different behavior for the J -test statistic. So exactly in what sense is $E[f(v_t, \theta)] = 0$ false? For now we'll consider **fixed mis-specification**. Specifically we'll suppose that

$$E[f(v_t, \theta)] = \mu(\theta), \quad \|\mu(\theta)\| > 0 \quad \forall \theta \in \Theta$$

Note that this situation can only occur if $q > p$ since we can always solve the population moment conditions *exactly* for θ in the just-identified case. Let \hat{S} be an estimator of the variance matrix of the moment conditions and let W be the probability limit of \hat{S}^{-1} . Then, if $\mu_* = \mu(\theta^*)$ is the probability limit of $\bar{g}_T(\hat{\theta})$, where θ^* is the solution to the projected moment conditions given by

the identifying restrictions, we have

$$\frac{1}{T}J_T = \bar{g}_T(\hat{\theta}_T)' \hat{S}^{-1} \bar{g}_T(\hat{\theta}_T) = \mu_*' W \mu_* + o_p(1)$$

As long as W is positive definite, $\mu_*' W \mu_* > 0$ since $\mu(\theta) > 0$ for all $\theta \in \Theta$. Thus, $J_T = T\mu_*' W \mu_* + o_p(T)$. In other words, under fixed mis-specification the J -test statistic *diverges at rate* T .

2 Andrews' GMM Moment Selection Criteria

The consistency and asymptotic normality results for GMM estimation rely on the assumption that the moment conditions used in estimation are correct. That is, they assume that $E[f(v_t, \theta_0)] = 0$. But what if we are unsure of this assumption? In many real-world applications we have a fairly large collection of moment functions, the q elements of f , some of which may have been derived under different economic or statistical assumptions than others. It could easily be the case that only *some* of the moment functions in f satisfy the moment conditions, while others do not. To take a simple example, we may have a collection of instrumental variables that arise from different sources or different assumptions on the DGP. Perhaps only some of these instruments are truly exogenous but we are unsure which. Andrews (1999) proposes a family of *moment selection criteria* (MSC) for this situation, in which the aim is to consistently select *any and all* elements of f that satisfy the moment condition, and eliminate those that do not.

Roughly speaking, the intuition is as follows. When we studied AIC, BIC and friends, we discussed how the maximized log-likelihood measures model fit but unfairly advantages models with more parameters. The various model selection criteria we examined amounted to adding some kind of “penalty” term to correct for this by *penalizing* more complicated models. In a similar vein, so long as we have more moment conditions than parameters, the J -test statistic provides a measure of how well the data “fit” the moment conditions:

the bigger the statistic, the greater the evidence that the moment conditions are violated. The problem is that J -test statistic tends to increase as we add additional moment conditions *even if they are correct*. Thus, if we simply compared J -statistics, we would be led to select *too few* moment conditions. To correct for this, Andrews (1999) considers a variety of “bonus terms” that *reward* estimators based on a larger number of moment conditions. Using this idea, he derives GMM analogues of AIC, BIC and the Hannan-Quinn information criterion, and studies the conditions under which a bonus term will yield consistent moment selection.

2.1 Notation

Let f_{max} be a $(q \times 1)$ vector containing all of the moment functions under consideration. Let c be a *selection vector*, a $(q \times 1)$ vector of ones and zeros indicating which elements of f_{max} we use in estimation for a *particular candidate specification*. Let \mathcal{C} denote the set of all candidates and $|c|$ denote the number of moment conditions used to estimate candidate c . Naturally, we require that there are at least as many moment conditions as parameters to estimate. Let $\hat{\theta}_T(c)$ be the efficient two-step GMM estimator based on the moment conditions $E[f(v_t, \theta, c)] = 0$ and define

$$\begin{aligned} V_\theta(c) &= [G_0(c)S(c)^{-1}G_0(c)]^{-1} \\ G_0(c) &= E[\nabla'_\theta f(v_t, \theta_0; c)] \\ S(c) &= \lim_{T \rightarrow \infty} Var \left[\frac{1}{\sqrt{T}} \sum_{t=1}^T f(v_t, \theta_0; c) \right] \\ J_T(c) &= T \bar{g}_T \left(\hat{\theta}_T(c); c \right)' \hat{S}_T(c)^{-1} \bar{g}_T \left(\hat{\theta}_T(c); c \right) \end{aligned}$$

where $\hat{S}(c)$ is a consistent estimator of $S(c)$.

2.2 Moment Selection Heuristics

So how should we choose c ? Two principles come to mind. First, we know that using only correctly specified moment conditions in estimation ensures that $\hat{\theta}_T \xrightarrow{p} \theta_0$. Thus, to ensure consistent estimation, we should seek to eliminate any moment conditions whose expectation is non-zero. Second, it can be shown (see Hall, 2005; Theorem 6.1) that adding additional correctly specified moment conditions *cannot increase* the asymptotic variance of our estimator. Putting these two pieces together, Andrews (1999) suggests that we attempt to identify the *maximal set of correctly specified moment conditions*.

But what exactly does this mean? Identification is a bit tricky when we start to consider the possibility that some of our moment conditions do not have expectation zero. The potential problem is that different subsets of f_{\max} could satisfy the population moment condition at *different values* of θ . We will need to rule this possibility out somehow. Let \mathcal{Z}^0 denote the set of all candidates c such that $E[f(v_t, \theta; c)] = 0$ for *some* $\theta \in \Theta$. Then, of all the candidates $c \in \mathcal{Z}^0$, let \mathcal{MZ}^0 denote those that contain the *maximum number* of elements of f_{\max} . For Andrews' suggestion to be meaningful, we need to assume that \mathcal{MZ}^0 contains *exactly one element*, which we'll call c_0 .

Andrews proposes adding a “bonus term” to the J -test statistic, leading to moment selection criteria (MSC) of the form

$$MSC(c) = J_T(c) - B(T, |c|)$$

where B is a “bonus term” that “rewards” specifications that use more moment conditions in estimation and may depend on sample size. In calculating the J -test statistic, Andrews recommends using a *centered* covariance matrix estimator

$$\hat{S}(c) = \frac{1}{T} \sum_{t=1}^T \left[f(v_t, \hat{\theta}_T(c); c) - \bar{g}_T(\hat{\theta}_T(c); c) \right] \left[f(v_t, \hat{\theta}_T(c); c) - \bar{g}_T(\hat{\theta}_T(c); c) \right]'$$

based on the weighing matrix that *would be* efficient if the moment conditions were correctly specified. This estimator is consistent for $S(c)$ *regardless* of whether the population moment conditions hold. To carry out moment selection, we choose c to *minimize* the criterion, defining $\hat{c}_T = \arg \min_{c \in \mathcal{C}} \text{MSC}(c)$.

2.3 Consistent Selection

The main point of Andrews (1999) is to establish sufficient conditions on the bonus term that guarantee consistent selection of any and all correctly specified moment conditions with probability approaching one in the limit. First we'll take a look at the conditions, and then the proof.

Regularity Conditions for the J -test Statistic

- (i) If $E[f(v_t, \theta; c)] = 0$ for a unique $\theta \in \Theta$, then $J_T(c) \xrightarrow{d} \chi^2_{|c|-p}$
- (ii) If $E[f(v_t, \theta; c)] \neq 0$ for a *all* $\theta \in \Theta$ then $T^{-1}J_T(c) \xrightarrow{p} a(c)$, a finite, positive constant that may depend on c .

Regularity Conditions for Bonus Term The bonus term can be written as $B(|c|, T) = \kappa_T h(|c|)$, where

- (i) $h(\cdot)$ is strictly increasing
- (ii) $\kappa_T \rightarrow \infty$ as $T \rightarrow \infty$ and $\kappa_T = o(T)$

Identification Conditions

- (i) $\mathcal{M}\mathcal{Z}^0 = \{c_0\}$
- (ii) $E[f(v_t, \theta_0; c_0)] = 0$ and $E[f(v_t, \theta; c_0)] \neq 0$ for any $\theta \neq \theta_0$

Theorem 2.1. *Under the preceding assumptions, $\hat{c}_T \xrightarrow{p} c_0$.*

Proof. We're trying to show that the moment conditions \widehat{c}_T selected by our criterion are consistent for the maximal set c_0 of correct moment conditions. By definition $\widehat{c}_T = \arg \min_{c \in \mathcal{C}} MSC_T(c)$, so we need to show that

$$\lim_{T \rightarrow \infty} P[\{MSC_T(c) - MSC_T(c_0) > 0, \forall c \neq c_0\}] = 1$$

To simplify the notation, define

$$\begin{aligned} \Delta_T(c, c_0) &= MSC_T(c) - MSC_T(c_0) \\ &= [J_T(c) - h(|c|)\kappa_T] - [J_T(c_0) - h(|c_0|)\kappa_T] \\ &= [J_T(c) - J_T(c_0)] + \kappa_T [h(|c_0|) - h(|c|)] \end{aligned}$$

Now, we are interested in $\Delta_T(c, c_0)$ *only* for situations in which $c \neq c_0$. Subject to this restriction, there are two cases, which we consider in turn.

Case 1 Consider $c_1 \neq c_0$ such that $E[f(v_t, \theta_1; c_1)] = 0$ for a *unique* θ_1 . In this case the first Regularity Condition for the J -test Statistic applies to *both* c_1 and c_0 and we have

$$J_T(c_1) - J_T(c_0) \xrightarrow{d} \chi^2_{|c_1|-p} - \chi^2_{|c_0|-p} = O_p(1)$$

By the first Identification Condition, c_0 is the *unique* maximal set of correct moment conditions. Hence $|c_0| > |c_1|$. Now, by the first Regularity Condition for the Bonus Term, h is strictly increasing. It follows that $h(|c_0|) - h(|c_1|) > 0$. By the second Regularity Condition for the Bonus Term, $\kappa_T \rightarrow \infty$. Thus,

$$\kappa_T [h(|c_0|) - h(|c|)] \rightarrow \infty$$

It follows that $\Delta_T(c_1, c_0) \rightarrow \infty$ and we obtain our desired result.

Case 2 Consider $c_2 \neq c_0$ such that $E[f(v_t, \theta; c_2)] \neq 0$ for any $\theta \in \Theta$. In this case, the *first* Regularity Condition for the J -test Statistic applies to c_0 , while

the *second* applies to c_2 so we have

$$T^{-1} [J_T(c_2) - J_T(c_0)] = a(c_2) + o_p(1) - T^{-1}O_p(1)$$

Now, whatever the value $[h(|c_0|) - h(|c|)]$ happens to be, it is definitely finite since h is strictly increasing by the first Regularity Condition for the Bonus Term, and both $|c|$ and $|c_0|$ are finite. By the second Regularity Condition for the Bonus Term, $\kappa_T = o(T)$. Hence,

$$T^{-1}\kappa_T [h(|c_0|) - h(|c|)] = o(1)$$

Putting the pieces together, we have

$$\begin{aligned} T^{-1}\Delta_T(c_2, c_0) &= a(c_2) + o_p(1) - T^{-1}O_p(1) + o(1) \\ &= a(c_2) + o_p(1) \end{aligned}$$

By the second Regularity Condition for the J -test Statistic, $a(c_2) > 0$. Thus, $T^{-1}\Delta_T(c_2, c_0) > 0$ with probability approaching one as $T \rightarrow \infty$. It follows that $\Delta_T(c_2, c_0) \rightarrow \infty$ with probability approaching one as $T \rightarrow \infty$, as required. \square

2.4 Which Criteria Are Consistent?

Among some other possibility, Andrews (1999) considers the following criteria which are constructed by making the bonus term resemble some of our old friends from maximum likelihood model selection:

$$\begin{aligned} \text{GMM-BIC}(c) &= J_T(c) - (|c| - p) \log(T) \\ \text{GMM-HQ}(c) &= J_T(c) - 2.01 (|c| - p) \log(\log(T)) \\ \text{GMM-AIC}(c) &= J_T(c) - 2 (|c| - p) \log(T) \end{aligned}$$

We see immediately that GMM-AIC does *not* satisfy the necessary conditions for consistency, since $\kappa_T = 2$ does not diverge as $T \rightarrow \infty$. In contrast, both

the GMM-BIC and GMM-HQ diverge as $T \rightarrow \infty$, so we simply need to check the requirement that $\kappa_T = o(T)$. For GMM-BIC we have

$$\lim_{T \rightarrow \infty} \frac{\log T}{T} = \lim_{T \rightarrow \infty} \frac{1}{T} = 0$$

by l'Hôpital's rule, and similarly for GMM-HQ

$$\lim_{T \rightarrow \infty} \frac{\log \log T}{T} = \lim_{T \rightarrow \infty} \frac{1}{\log T} = 0$$

Thus both GMM-BIC and GMM-HQ provide consistent moment selection.

2.5 Asymptotics for GMM-AIC

We saw in the previous subsection that GMM-AIC does not satisfy the sufficient conditions for consistent moment selection. The question remains: how does this criterion behave in the limit? To answer this question, we revisit the proof of consistent selection from above. It turns out that GMM-AIC behaves *differently* in the two cases considered in the proof. Combining them, we will see that GMM-AIC is *not* a consistent moment selection criterion.

Case 2 In this case, we examined $c_2 \neq c_0$ such that $E[f(v_t, \theta; c_2)] \neq 0$ for any $\theta \in \Theta$. In other words, the moment conditions indexed by c_2 are *not* satisfied for *any* parameter value θ . Asymptotically, GMM-AIC will *never* select such a set of moment conditions. To see why, recall that $\kappa_T = 2$ for GMM-AIC. Although it does not diverge, this choice of κ_T is *still* $o(T)$. Thus, the argument from Case 2 *still applies* to the GMM-AIC. We did not in fact use the assumption that κ_T diverges in the proof of this case!

Case 1 In this case, we examined $c_1 \neq c_0$ such that $E[f(v_t, \theta_1; c_1)] = 0$ for a *unique* θ_1 . In other words, we considered a situation in which there *is* a parameter vector θ_1 at which the moment conditions indexed by c_1 are satisfied. Now, the difference of J -test statistics continues to be $O_p(1)$ regardless

of the choice of κ_T , provided the regularity conditions are satisfied. Thus, substituting $\kappa_T = 2$, we have

$$\Delta_T(c_1, c_0) = O_p(1) + 2[h(|c_0|) - h(|c|)]$$

But since the second term is a *constant*, this is simply $\Delta_T(c_1, c_0) = O_p(1)$. In other words, the GMM-AIC is a *random variable*, even in the limit as $T \rightarrow \infty$.

So where does this leave us? In Case 2 GMM-AIC consistently selects c_0 , but in Case 1 GMM-AIC is *random even in the limit*. Putting these two results together, we see that, although it will never select a set of false moment conditions, GMM-AIC chooses *randomly* among the set of correct moment conditions. In other words, it will not necessarily select c_0 as $T \rightarrow \infty$.

2.6 Extensions of Andrews (1999)

Two followup papers extended the criteria described above. Andrews and Lu (2001) consider simultaneous model *and* moment selection for GMM. The basic idea is the same, except that the parameter vector θ is restricted under some specifications. For example, we may consider setting a coefficient to zero. Accordingly, the “bonus term” depends both on the number of moment conditions used in estimation and the number of parameters that are estimated. Hong, Preston & Shum (2003) extend Andrews and Lu (2001) to Generalized Empirical Likelihood estimators. For details on this class of estimators and their properties, see Newey & Smith (2004).

2.7 Drawbacks to Andrews’ Approach

Andrews (1999) has a very specific goal: to state conditions under which it is possible to carry out consistent moment selection for GMM. This is an important and very useful contribution. Nevertheless, there are several reasons why the framework used in Andrews (1999) may not be appropriate in practical applications of GMM moment selection.

First, the identification condition $\mathcal{M}\mathcal{Z}^0 = \{c_0\}$ is stronger than it may appear. Section 7.3.1 of Hall (2005) gives an example: a linear IV model with one endogenous regressor, jointly normal errors, and eight instruments. Six of the instruments are exogenous, but two are in fact endogenous. The goal of moment selection is to find the exogenous instruments. Even in this very simple setting, the identification condition fails: there are two *different* candidates, each containing six moment conditions, for which it is possible to find a parameter value at which the population moment conditions are satisfied. One of these parameters is the true θ and the other is not. The problem with the identification assumption isn't so much that it's strong. Without strong assumptions, it's hard to learn anything. The problem is that it's not especially *transparent*: when considering a particular problem it can be hard to get a handle on whether this assumption makes sense.

A second problem concerns irrelevant moment conditions. The idea of using any and all correctly specified moment conditions in estimation is based on the fact that the asymptotic variance of the GMM estimator *cannot increase* as we use additional moment conditions in estimation. The finite-sample situation, however, can be very different. Moment conditions that add very little information, so-called “irrelevant moment conditions,” can lead to very poor finite sample performance. The GMM-MSR of Andrews (1999) does not address this problem. Two papers that do are Hall & Peixe (2003) and Hall, Inoue, Jana & Shin (2003). More recently, Chen & Liao (2013) suggest using LASSO, which we'll study in an upcoming lecture, to choose the valid and relevant moment conditions.

A third issue concerns the nature of the analogy between the AIC and BIC and their GMM-MSR counterparts. Like BIC, GMM-BIC is consistent and like AIC, GMM-AIC is not. The relationship ends here, however. As we saw in our first lectures of the semester there is a *very specific idea* behind both the AIC and the BIC: the former attempts to correct the bias in the maximized log-likelihood as an estimator of the KL divergence, and the latter provides a large-sample approximation to the Bayesian posterior model probabilities under a

uniform prior. Neither of these ideas has anything to do with the arguments behind the GMM-MSR criteria. Beyond conditions on the asymptotic behavior of the bonus term, any relationship to the AIC and BIC is merely cosmetic. This raises an interesting question: can we re-work any of the principles we used to derive model selection criteria for maximum likelihood so that they can be applied to GMM moment selection? The answer turns out to be *yes*, as we will see below.

3 The Focused Moment Selection Criterion

This section is based on my working paper “Using Invalid Instruments on Purpose: Focused Moment Selection for GMM.” For the current version, see my website: <http://www.ditraglia.com>.¹

3.1 Introduction

In practical applications of GMM, we are rarely interested in determining which moment conditions are correct. More commonly, our goal is to answer a *research question*, typically involving a parameter of interest μ that depends on the underlying GMM parameter vector θ . Accordingly, it might make sense to try to get “good” estimates of μ , *regardless* of whether this involves using correct or incorrect moment conditions. The basic idea is that we might want to use a moment condition that is *slightly mis-specified* provided that it is sufficiently informative about μ : the decrease in variance could easily outweigh the increase in bias in a MSE-sense. This is very similar to the idea that underlies Mallows’s C_p , Akaike’s Final Prediction Error, and, you guessed it, the Focused Information Criterion of Hjort & Claeskens (2003). My approach is most similar to the FIC, so I’ve named it the Focused Moment Selection Criterion, or FMSC for short.

¹If you’re really interested in this, check back again after the semester is over. I’ll be posting an updated version with R and C++ source code soon.

3.2 Overview of FMSC Derivation

Local Mis-specification

$$E \begin{bmatrix} g(Z_{ni}, \theta_0) \\ h(Z_{ni}, \theta_0) \end{bmatrix} = \begin{bmatrix} 0 \\ \tau/\sqrt{n} \end{bmatrix}$$

Identification Condition $E[g(Z_{ni}, \theta_0)] = 0$ identifies θ_0 .

Moment Selection Matrix The matrix of zeros and ones Ξ_S selects the moment conditions used to estimate candidate specification S .

Candidate GMM Estimator

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

where

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

and \widetilde{W} is positive semi-definite weighting matrix that converges in probability to a positive definite matrix W .

Some Notation Let Z denote the limiting random variable, for which *all moment conditions are correctly specified* and define

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

and

$$\Omega = Var[f(Z, \theta_0)] = \begin{bmatrix} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{bmatrix}$$

High-Level Condition I – Expansion for GMM

$$\sqrt{n}(\hat{\theta}_S - \theta_0) = -K_S \sqrt{n}[\Xi_S f_n(\theta_0)] + o_p(1)$$

Where

$$\begin{aligned} F_S &= \Xi_S F \\ W_S &= \Xi_S W \Xi_S' \\ M_S &= \Xi_S M \\ \Omega_S &= \Xi_S \Omega \Xi_S' \\ K_S &= [F_S' W_S F_S]^{-1} F_S' W_S \end{aligned}$$

High-Level Condition II – CLT $\sqrt{n}f_n(\theta_0) \xrightarrow{d} M$ where

$$M = \begin{bmatrix} M_g \\ M_h \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \tau \end{bmatrix}, \Omega\right)$$

Asymptotic Distribution of GMM Estimator $\sqrt{n}(\hat{\theta}_S - \theta_0) \rightarrow_d -K_S M_S$

Target Parameter $\mu_0 = \mu(\theta_0)$, $\hat{\mu}_S = \mu(\hat{\theta}_S)$ where $\mu(\cdot)$ is differentiable.

Asymptotic Distribution of $\hat{\mu}_S$

$$\sqrt{n}(\hat{\mu}_S - \mu_0) \rightarrow_d -\nabla_{\theta}\mu(\theta_0)' K_S M_S$$

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

Asymptotically Unbiased Estimator of τ Here is where we use the identification condition. Let $\hat{\theta}_{valid}$ denote the *valid estimator*, that is the estimator that uses only moment conditions in g . Our identification assumption ensures

that this estimator identifies θ_0 . We have

$$\begin{aligned}\sqrt{n} \left(\hat{\theta}_{valid} - \theta_0 \right) &= -K_v \sqrt{n} g_n(\theta_0) + o_p(1) \\ &\xrightarrow{d} -K_v M_h\end{aligned}$$

So this estimator has *no asymptotic bias*. Under local mis-specification, no consistent estimator of τ exists, but we can construct an asymptotically unbiased estimator by plugging $\hat{\theta}_{valid}$ into the sample analogue of the h moment conditions. In particular, the estimator is $\sqrt{n} h_n(\hat{\theta}_{valid})$. By a Mean-Value Expansion:

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

Thus, $\hat{\tau} \xrightarrow{d} \Psi M$ where $\Psi = \begin{bmatrix} -H K_v & \mathbf{I}_q \end{bmatrix}$, so we have $\Psi M \sim \mathcal{N}_q(\tau, \Psi \Omega \Psi')$.

Asymptotically Unbiased Estimator of $\tau \tau'$ Let $\hat{\Omega}$ and $\hat{\Psi}$ be consistent estimators of Ω and Ψ . Then, $\hat{\tau} \hat{\tau}' - \hat{\Psi} \hat{\Omega} \hat{\Psi}' \rightarrow_d \Psi (M M' - \Omega) \Psi'$. That is, $\hat{\tau} \hat{\tau}' - \hat{\Psi} \hat{\Omega} \hat{\Psi}'$ provides an asymptotically unbiased estimator of $\tau \tau'$.

The Focused Moment Selection Criterion At long last, we can write down the FMSC. The following expression provides an asymptotically unbiased estimator of $\text{AMSE}(\hat{\mu}_S)$

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

3.3 A Very Simple Example

This is the simplest possible example of the FMSC: choosing between OLS and 2SLS. Suppose that we have a vector of valid instruments \mathbf{z} and we want to estimate the coefficient β on a single endogenous regressor x in the following linear system:

$$\begin{aligned} y_i &= \beta x_i + \epsilon_i \\ x_i &= \mathbf{z}_i' \boldsymbol{\pi} + v_i \end{aligned}$$

It's no problem accomodating additional exogenous control regressors: just project them out of the system before proceeding.

Now, if we want to estimate β , one option is to use the 2SLS estimator

$$\tilde{\beta} = [\mathbf{x}' P_Z \mathbf{x}]^{-1} \mathbf{x}' P_Z \mathbf{y}$$

Another option is to use the OLS estimator

$$\hat{\beta} = (\mathbf{x}' \mathbf{x})^{-1} \mathbf{x}' \mathbf{y}$$

But why on earth would we ever want to use OLS? If x is endogenous and we have some valid instruments, shouldn't we use 2SLS? The answer, as you may have guessed is: "it depends: there's a bias-variance tradeoff." By using 2SLS, we guarantee that our estimator will be asymptotically unbiased, but this comes at the cost of a much higher asymptotic variance. If x is not *too endogenous* it could make sense to use OLS rather than IV. This is exactly the idea that the FMSC tries to capture.

To put this problem into the FMSC framework, we write

$$E_n \begin{bmatrix} \mathbf{z}_i \epsilon_i \\ x_i \epsilon_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \tau / \sqrt{n} \end{bmatrix}$$

Because everything is linear, it's straightforward to derive the limiting distri-

butions of the OLS and 2SLS estimators. After some algebra, we find that the AMSE expressions take a very simple form:

$$\begin{aligned}\text{AMSE(OLS)} &= \frac{\tau^2}{\sigma_x^4} + \frac{\sigma_\epsilon^2}{\sigma_x^2} \\ \text{AMSE(2SLS)} &= \frac{\sigma_\epsilon^2}{\gamma^2}\end{aligned}$$

where $\sigma_x^2 = \gamma^2 + \sigma_v^2$, $\gamma^2 = \boldsymbol{\pi}' Q_z \boldsymbol{\pi}$, and $Q_z = \text{plim } Z'Z/n$. Thus, the AMSE of the OLS estimator is lower than that of the IV estimator whenever

$$\frac{\tau^2}{\sigma_v^2 \sigma_\epsilon^2} < \frac{\sigma_x^2}{\gamma^2}$$

The usual estimators of σ_x^2 , γ^2 , and σ_v^2 remain consistent under local mis-specification:

$$\begin{aligned}\hat{\sigma}_x^2 &= n^{-1} \mathbf{x}' \mathbf{x} \xrightarrow{p} \boldsymbol{\pi}' Q_z \boldsymbol{\pi} + \sigma_v^2 \\ \hat{\gamma}^2 &= n^{-1} \mathbf{x}' Z (Z' Z)^{-1} Z' \mathbf{x} \xrightarrow{p} \boldsymbol{\pi}' Q_z \boldsymbol{\pi} \\ \hat{\sigma}_v^2 &= \hat{\sigma}_x^2 - \hat{\gamma}^2\end{aligned}$$

To get a consistent estimator of σ_ϵ^2 under local mis-specification, we can use either the residuals from OLS or 2SLS, but 2SLS may be more robust. To implement the FMSC for this problem, we simply need an asymptotically unbiased estimator of τ^2 .

The asymptotically unbiased estimator of τ for this problem is

$$\hat{\tau} = \sqrt{n} \left[\mathbf{x}' (\mathbf{y} - \mathbf{x} \tilde{\beta}) / n \right] = n^{-1/2} \mathbf{x}' (\mathbf{y} - \mathbf{x} \tilde{\beta})$$

since $\hat{\tau} \xrightarrow{d} N(\tau, V)$ where

$$V = \sigma_\epsilon^2 \sigma_x^2 \left(\frac{\sigma_v^2}{\gamma^2} \right)$$

Hence

$$\hat{\tau}^2 - \hat{\sigma}_\epsilon^2 \hat{\sigma}_x^2 \left(\frac{\hat{\sigma}_v^2}{\hat{\gamma}^2} \right)$$

is an asymptotically unbiased estimator of τ^2 . Substituting this quantity and rearranging, the FMSC tells us to use the OLS estimator whenever

$$\hat{T}_{FMSC} = \frac{\hat{\tau}^2 \hat{\gamma}^2}{\hat{\sigma}_v^2 \hat{\sigma}_\epsilon^2 \hat{\sigma}_x^2} < 2$$

After some algebra, it turns out that \hat{T}_{FMSC} is *numerically equivalent* to the Hausman Test statistic and that $\hat{T}_{FMSC} \xrightarrow{d} \chi^2(1)$ when $\tau = 0$. Thus, in the simple example of choosing between OLS and 2SLS, the FMSC is identical to carrying out a Hausman Test with a critical value of 2, which corresponds to a 16% significance level. Notice that this is exactly the same significance level that appeared when we interpreted the AIC as a hypothesis test in our simple example of estimating a normal mean! This relationship only holds, so far as I know, for the present example. Viewed from the opposite perspective, these derivations indicate that the textbook procedure of using a Hausman Test to choose between OLS and 2SLS can be rigorously grounded in a loss-based framework. The usual significance levels of 5 or 10%, however, are *too lenient*: they would lead us to use OLS in some situations that do *not* lead to a favorable bias-variance tradeoff.