COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1146R

CONSISTENT MOMENT SELECTION PROCEDURES FOR GENERALIZED METHOD OF MOMENTS ESTIMATION

Donald W. K. Andrews

January 1997
Revised November 1997

Consistent Moment Selection Procedures for Generalized Method of Moments Estimation

Donald W. K. Andrews[1]
Cowles Foundation for Research in Economics
Yale University

November 1996
Revised: October 1997

Abstract

This paper considers a generalized method of moments (GMM) estimation problem in which one has a vector of moment conditions, some of which are correct and some incorrect. The paper introduces several procedures for consistently selecting the correct moment conditions. The procedures also can consistently determine whether there is a sufficient number of correct moment conditions to identify the unknown parameters of interest.

The paper specifies moment selection criteria that are GMM analogues of the widely used BIC and AIC model selection criteria. (The latter is not consistent.) The paper also considers downward and upward testing procedures.

All of the moment selection procedures discussed in the paper are based on the minimized values of the GMM criterion function for different vectors of moment conditions. The procedures are applicable in time series and cross-sectional contexts.

Application of the results of the paper to instrumental variables estimation problems yields consistent procedures for selecting instrumental variables.

Keywords: Akaike information criterion, Bayesian information criterion, consistent selection procedure, downward testing procedure, generalized method of moments estimator, instrumental variables estimator, model selection, moment selection, test of over-identifying restrictions, upward testing procedure.

JEL Classification: C12, C13, C52.

1. Introduction

Empirical researchers using generalized method of moments (GMM) estimation methods often find that the J test of over-identifying restrictions rejects the null hypothesis. Rejection of the null indicates that not all moment conditions are correct. In such cases, it may be useful to employ a moment selection procedure that estimates which moments are correct and which are incorrect.

This paper introduces several such moment selection procedures. First, we consider procedures based on a moment selection criterion (MSC). Minimization of the MSC over the parameter space $\mathcal{C}$ for the selection vectors yields an estimate of the correct moment conditions. We introduce GMM analogues of the widely used BIC, AIC, and HQIC model selection criteria. We refer to these criteria as GMM–BIC, GMM–AIC, and GMM–HQIC. These criteria are based on the J test statistic for over-identifying restrictions. A "bonus" term is subtracted from the J test statistic that rewards selection vectors that employ more moment conditions. We demonstrate that the GMM–BIC, GMM–AIC, and GMM–HQIC moment selection criteria are the proper analogues of BIC, AIC, and HQIC by showing that they use the same asymptotic trade-off between the "model fit" and the "number of parameters."

For specificity, we define the GMM–BIC criterion here. Let c denote a moment selection vector, i.e., a vector that selects some moment conditions but not others. Let $|c|$ denote the number of moment conditions selected by c. Let $J_n(c)$ denote the J test statistic for over-identifying restrictions constructed using the moment selection vector c. Let $\mathcal{C}$ be the parameter space for the moment selection vector. Let p be the dimension of the vector θ to be estimated by GMM. Let n denote the sample size. Then, the GMM-BIC moment selection criterion chooses the vector in $\mathcal{C}$ that minimizes

$$J_n(c) - (|c| - p) \ln n \text{ over } \mathcal{C}. \tag{1.1}$$

Note that $|c| - p$ is the number of over-identifying restrictions.

We also consider two testing procedures that can be used to select correct moment conditions. These procedures are similar to informal methods based on the J test often employed by empirical researchers to determine which moments to use. We consider downward testing (DT) and upward testing (UT) procedures. Both are based on the J test.

We specify conditions under which the MSC and testing procedures are consistent. The conditions allow for independent identically distributed (iid), stationary and ergodic, independent non-identically distributed (inid), and strong mixing non-identically distributed (mnid) random variables (rv's). Under somewhat stronger conditions, the procedures are strongly consistent, see Andrews (1997b). The GMM–AIC procedure does not satisfy the conditions for consistency or strong consistency, but the other procedures do.

Our results cover the case of linear instrumental variables (IV) estimation as a special case. Thus, the procedures introduced in the paper can be used to consistently select IVs for two stage least squares (2SLS) or two stage instrumental variables (2SIV) estimators (see White (1982) regarding the latter).

The finite sample behavior of the moment selection procedures is investigated in Andrews (1997b) via a Monte Carlo experiment. The GMM–BIC, DT, and UT procedures perform best and about equally well in the experiment. The GMM–HQIC procedure is next best and the GMM–AIC procedure is worst overall.

The Monte Carlo results indicate that moment selection procedures may be useful tools for empirical researchers if used prudently. In particular, the use of a small parameter space for the selection vectors seems highly desirable. If one uses the procedures indiscriminately, with a large parameter space, the procedures may perform poorly.

We find that a simple method can be used to detect whether a MSC is reliable. In those cases where an MSC performs poorly, there are typically two or more selection vectors that yield MSC values that are close to the minimum and that yield parameter estimates that differ noticeably from each other. In cases where a moment selection procedure performs well, the latter typically does not occur. Thus, one can use this as a condition for detecting reliability of the MSC.

Topics for future research include the following: determination of optimality properties for some moment selection procedure, investigation of the use of the bootstrap to assess the performance of moment selection procedures and/or to improve the performance of moment selection procedures, extension of the results to cover simultaneous moment and model selection, see Lu and Andrews (1997), and analysis of the asymptotic behavior of the procedures in the context of weak instruments, as in Stock and Wright (1997).

We now discuss the literature that is related to the procedures introduced in this paper. Gallant and Tauchen (1996) recently address the issue of selecting a small number of efficient moments from a large pool of correct moments. This is a different problem from that addressed here. Gallant, Hsieh, and Tauchen (1997) consider using t-ratios for individual moment conditions as diagnostics for moment failure. Their t-ratios cannot be used to construct consistent moment selection procedures and their use as diagnostics for moment failure is questionable, because the inclusion of any incorrect moments typically yields an inconsistent parameter estimator, which in turn leads to rejection of all moments asymptotically, not just incorrect moments. Kolaczyk (1995) considers an analogue of the AIC in an empirical likelihood context, but his analogue is a model selection criterion not a moment selection criterion. The closest results in the literature to those given here seem to be Eichenbaum, Hansen, and Singleton's (1988, Appendix C) test of whether a given subset of moment conditions is correct or not. They propose a likelihood ratio-like test based on the GMM objective function for a single block of potentially incorrect moments. They do not consider moment selection criteria, such as GMM-BIC. Somewhat related to the procedures considered here are the results of Smith (1992) and Pesaran and Smith (1994).

In terms of the methods used, the model selection literature is the closest literature to the present paper. We borrow from this literature extensively in the present paper. Much of this literature has focussed on the lag length selection problem for autoregressive and autoregressive-moving average (ARMA) models and, more

generally, the regressor selection problem for regression models. The BIC criterion is introduced by Schwarz (1978), Rissanen (1978), and Akaike (1977); the AIC criterion by Akaike (1969); and the HQIC criterion by Hannan and Quinn (1979). Model selection via upward Lagrange multiplier testing is introduced by Pötscher (1983) for the ARMA selection problem. Consistency, strong consistency, or lack thereof of these procedures are established by Shibata (1976), Hannan (1980, 1982), Hannan and Deistler (1988), and Pötscher (1989), as well as some of the references above. For the literature on regressor selection, see Amemiya (1980), Pötscher (1989), and references therein.

The remainder of this paper is organized as follows. Section 2 describes the moment selection problem and introduces definitions, notation, and assumptions that are used throughout the paper. Section 3 introduces a class of moment selection criteria, including the GMM–BIC, GMM–AIC, and GMM–HQIC criteria, and provides a condition under which such criteria are consistent. Sections 4 and 5 introduce downward and upward testing procedures, respectively, for selecting moments and provides conditions under which they are consistent. Section 6 establishes that GMM–BIC is the appropriate analogue of BIC etc. An Appendix of Proofs provides proofs of results stated in Sections 2–6.

2. Description of the Moment Selection Problem, Definitions, and Assumptions

2.1. The Moment Selection Problem

We have an infinite sequence of random variables $Z_1, ..., Z_n, ...$ drawn from an unknown probability distribution P^0 (the data generating process) that is assumed to belong to a class $\mathcal{P}$ of probability distributions. The class $\mathcal{P}$ allows for the cases where the random variables are iid, inid, stationary and ergodic, weakly dependent and non-identically distributed, etc.

We have a random vector of moment conditions

$$G_n(\theta) : \Theta \to R^r \tag{2.1}$$

and a random $r \times r$ weight matrix W_n, both of which depend on $\{Z_i : i \leq n\}$, and $\Theta \subset R^p$. Typically, the moment conditions are of the form $G_n(\theta) = \frac{1}{n}\sum_{i=1}^n m(Z_i, \theta)$.

We assume that $G_n(\theta)$ converges in probability as $n \to \infty$ to a function $G^0(\theta)$ $\forall \theta \in \Theta$, $\forall P^0 \in \mathcal{P}$. (A formal statement of assumptions is provided below.) Usually, this holds by a weak law of large numbers (LLN) and $G^0(\theta)$ is the expectation of $G_n(\theta)$ or its limit as $n \to \infty$. The superscript "0" on $G^0(\theta)$, and on various other quantities introduced below, denotes dependence on P^0.

In the standard GMM framework (which is not adopted here), one assumes that all r moment conditions are correct. That is, for some $\theta^0 \in \Theta$, one has $G^0(\theta^0) = \mathbf{0}$. Furthermore, to achieve identification, one assumes that θ^0 is the unique solution to these equations. The parameter θ^0 is then called the "true" value of θ. In this case,

the standard GMM estimator $\widehat{\theta}_n$ of θ^0 is defined to minimize

$$G_n(\theta)'W_nG_n(\theta) \text{ over } \theta \in \Theta.^2 \tag{2.2}$$

The GMM estimator $\widehat{\theta}_n$ is consistent for θ^0 under minimal (and well-known) additional assumptions.

Often in empirical applications, however, researchers find that the J test of overidentifying restrictions, see Hansen (1982), rejects the null hypothesis that all r moment conditions are correct. Thus, it seems useful to consider statistical inference in the case where not all of the moment conditions are correct. That is what we do here. We presume that the researcher does not know a priori which moment conditions are correct. (Otherwise he would discard the incorrect moment conditions and be faced with the standard situation considered in the literature.)

Below we show that under certain assumptions it is possible to consistently estimate which moment conditions are correct and which are incorrect, given a suitable definition of "correct." This allows one to construct a GMM estimator that relies only on correct moment conditions asymptotically, provided there is a sufficient number of them.

2.2. Definition of the Correct Selection Vector

Here, we define the vector of "correct" moment conditions. We let $c \in R^r$ denote a *moment selection vector*. By definition, c is a vector of zeros and ones. If the j-th element of c is a one, then the j-th moment condition is included. If the j-th element is a zero, then it is not included. Let

$$\mathcal{S} = \{c \in R^r : c_j = 0 \text{ or } 1 \quad \forall 1 \leq j \leq r, \text{ where } c = (c_1, ..., c_r)'\}. \tag{2.3}$$

Let $|c|$ denote the number of moments selected by c, i.e., $|c| = \sum_{j=1}^r c_j$ for $c \in \mathcal{S}$. For any r-vector v and any $c \in \mathcal{S}$ with $c \neq \mathbf{0}$, let v_c denote the $|c|$-vector that results from deleting all elements of v whose coordinates equal coordinates of elements of c that are zeros. Thus, $G_{nc}(\theta)$ is the $|c|$-vector of moment conditions that are specified by $c \in \mathcal{S}$. For $c = \mathbf{0}$, let $v_{|c|} = 0$ $(\in R)$.

We now define the "correct" selection vector c^0 of moment conditions. Let $c^0(\theta)$ be the r vector of zeros and ones whose j-th element is one if the j-th element of $G^0(\theta)$ equals zero and is zero otherwise. Thus, $c^0(\theta)$ indicates which moments equal zero asymptotically when evaluated at the parameter vector θ. Define

$$\mathcal{Z}^0 = \{c \in \mathcal{S} : |c| = c'c^0(\theta) \text{ for some } \theta \in \Theta\}. \tag{2.4}$$

As defined, $\mathcal{Z}^0$ is the set of selection vectors in $\mathcal{S}$ that select *only* moment conditions that equal zero asymptotically for some $\theta \in \Theta$. (The notation "$\mathcal{Z}^0$" is meant to remind one of "zero under P^0".) Define

$$\mathcal{MZ}^0 = \{c \in \mathcal{Z}^0 : |c| \geq |c^*| \ \forall c^* \in \mathcal{Z}^0\}. \tag{2.5}$$

As defined, $\mathcal{MZ}^0$ is the set of selection vectors in $\mathcal{Z}^0$ that maximize the number of selected moments out of selection vectors in $\mathcal{Z}^0$. (The notation "$\mathcal{MZ}^0$" denotes "maximal zeros under P^0.")

For given $P^0 \in \mathcal{P}$, we consider the following assumption:

Assumption IDc^0. *$\mathcal{MZ}^0$ contains a single element c^0.*

When Assumption IDc^0 holds, we call c^0 the "correct" selection vector. The correct selection vector c^0 has the property that it uniquely selects the maximal number of moment conditions that equal zero asymptotically for some parameter $\theta \in \Theta$. Depending upon P^0, Assumption IDc^0 may or may not hold. Below we analyze the properties of moment selection procedures both when this identification assumption holds and when it fails to hold.

If the maximum number of moment conditions that are zero asymptotically is p or less, i.e., $|c| \leq p$ for $c \in \mathcal{MZ}^0$, then Assumption IDc^0 typically does not hold. The reason is that whenever there are as many or more parameters p as moment conditions $|c|$ there is usually some p-vector $\theta_c \in \Theta$ that solves the $|c|$ moment conditions $G_c(\theta) = \mathbf{0}$. Thus, Assumption IDc^0 typically requires one or more *over-identifying* restrictions for it to hold. That is, it requires $|c| > p$ for $c \in \mathcal{MZ}^0$.

Next, for distributions P^0 for which Assumption IDc^0 holds, we consider the following condition:

Assumption ID$\boldsymbol{\theta}^0$. $G^0_{c^0}(\theta) = \mathbf{0}$ has a unique solution $\theta^0 \in \Theta$.

When Assumption IDθ^0 holds, we call θ^0 the "true" value of θ. The true value θ^0 has the property that it sets the moment conditions selected by c^0 to be zero and is the unique parameter vector θ that does so.

Note that the standard GMM situation considered in the literature corresponds to the case where $\mathcal{MZ}^0 = \{\mathbf{1}_r\}$ and Assumption IDθ^0 holds, where $\mathbf{1}_r$ denotes an r-vector of ones. The former condition implies that Assumption IDc^0 holds.

To obtain consistent estimators of c^0 when Assumption IDc^0 holds, it turns out that one does not need Assumption IDθ^0 to hold. To obtain consistent estimators of both c^0 and θ^0, however, one needs both Assumptions IDc^0 and IDθ^0 to hold.

Next, we discuss Assumptions IDc^0 and IDθ^0 in the context of linear IV estimation. Consider the iid linear regression model $Y_i = X_i'\theta^* + U_i$ for $i = 1, ..., n$ under P^0, where $EU_i = 0$ and $E||X_i|| < \infty$. We consider the IVs $\widetilde{Z}_i \in R^r$, where $A^0 = E\widetilde{Z}_i X_i' \in R^{r \times p}$ and $\rho^0 = E\widetilde{Z}_i U_i \in R^r$. The moment conditions in this case are $G_n(\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\theta)\widetilde{Z}_i$ and the corresponding limit function is $G^0(\theta) = E(Y_i - X_i'\theta)\widetilde{Z}_i = \rho^0 - A^0(\theta - \theta^*)$. Let $c^* \in \mathcal{S}$ denote the selection vector that selects all of the IVs that are not correlated with the error U_i. Thus, the $j-$th element of c^* is one if the corresponding element of ρ^0 is zero and is zero otherwise. We assume that there are more good IVs than parameters, i.e., $|c^*| > p$. In this context, the parameter of interest is θ^* and the selection vector of correct IVs is c^*.

A question of interest is: When do Assumptions IDc^0 and IDθ^0 hold with $c^0 = c^*$ and $\theta^0 = \theta^*$? It is easy to see that $c^* \in \mathcal{Z}^0$. Let A^0_c denote the matrix A^0 with the rows corresponding to zeros in c deleted. Then, Assumption IDc^0 holds with $c^0 = c^*$ if and only if ρ^0_c is not in the column space of A^0_c for any $c \neq c^*$ with $|c| \geq |c^*|$, where $\rho^0_c \neq 0 \in R^{|c|}$, $A^0_c \in R^{|c| \times p}$, and $|c| > p$. Only very special A^0 and ρ^0 matrices violate this condition. If the former condition holds, then Assumption IDθ^0 holds with $\theta^0 = \theta^*$ if and only if $A^0_{c^*}$ is full column rank p, because $G^0_{c^*}(\theta) = A^0_{c^*}(\theta - \theta^*)$.

2.3. The J-test Statistic

All of the moment selection procedures considered below are based on the J test statistic used for testing over-identifying restrictions, see Hansen (1982). The J test statistic based on the vector of moment conditions selected by c is defined to be

$$J_n(c) = n \inf_{\theta \in \Theta} G_{nc}(\theta)' W_{nc} G_{nc}(\theta). \tag{2.6}$$

Here, W_{nc} is the $|c| \times |c|$ weight matrix employed with the moment conditions $G_{nc}(\theta)$. For example, W_{nc} might be defined such that it is an asymptotically optimal weight matrix when the moment conditions selected by c are correct.[3] (Note that W_{nc} is not necessarily equal to the matrix that results from deleting the rows and columns of W_n that correspond to the elements of c that are zeros.) By definition, when $c = \mathbf{0}$, $W_{nc} = 0$ $(\in R)$.

The GMM estimator based on the moment conditions $c \in \mathcal{C}$ is defined to be any vector $\widehat{\theta}_n(c) \in \Theta$ for which

$$G_{nc}(\widehat{\theta}_n(c))' W_{nc} G_{nc}(\widehat{\theta}_n(c)) = \inf_{\theta \in \Theta} G_{nc}(\theta)' W_{nc} G_{nc}(\theta). \tag{2.7}$$

Thus, the $J_n(c)$ test statistic also can be written as $J_n(c) = nG_{nc}(\widehat{\theta}_n(c))' W_{nc} G_{nc}(\widehat{\theta}_n(c))$.[4]

2.4. The Parameter Space for the Moment Selection Vectors

Next, we discuss the basis for various moment selection procedures. We consider estimation of c^0 via an estimator that we denote generically by $\hat{c}$. The parameter space for $\hat{c}$ is denoted by $\mathcal{C} \subset \mathcal{S}$. The parameter space $\mathcal{C}$ is defined to include $c = \mathbf{0}$.

The parameter space $\mathcal{C}$ should be a very much smaller set than $\mathcal{S}$ that exploits the nested or hierarchical structure that typically arises. Otherwise, the finite sample behavior of $\hat{c}$ will be poor and computation will be difficult. First, $\mathcal{C}$ should incorporate the information that certain moment conditions are assumed to be correct. Second, if applicable, $\mathcal{C}$ should incorporate information that certain blocks of moment conditions are either correct or incorrect block by block rather than moment condition by moment condition. This often occurs in the common case where the moment function $m(Z_i, \theta)$ is of the form

$$m(Z_i, \theta) = m^*(Z_i, \theta) \otimes X_i, \tag{2.8}$$

where $m^*(Z_i, \theta) \in R^{r_1}$, $X_i \in R^{r_2}$, X_i is a subvector of Z_i, and $r = r_1 \times r_2$. Depending upon an underlying model, one might assume that the moments $m^*(Z_i, \theta) X_{ij} \in R^{r_1}$ are all correct or all incorrect for a given $j \leq r_2$, where $X_i = (X_{i1}, ..., X_{ir_2})'$. This is plausible if a model implies that $Em^*(Z_i, \theta) X_{ij} = \mathbf{0}$ or not depending on whether X_{ij} is in the information set of decision maker i or of the decision maker at time i.

Third, if $X_{i\ell}$ is a function of X_{ij}, such as X_{ij}^2, then $\mathcal{C}$ should incorporate, when appropriate, the feature that the whole block of moments $m^*(Z_i, \theta) \otimes (X_{ij}, X_{i\ell})'$ is either correct or incorrect. Fourth, $\mathcal{C}$ should incorporate, again when appropriate, the feature that a whole block of moments $m_k^*(Z_i, \theta) \otimes X_i$ is either correct or incorrect for some $k \leq r_1$, where $m^*(Z_i, \theta) = (m_1^*(Z_i, \theta), \ ... \ , m_{r_1}^*(Z_i, \theta))'$.

2.5. Definitions of Consistency

We introduce two definitions of consistency. The first is the standard definition of consistency that is analogous to the definition used in the model selection literature. The second, called *s-consistency*, is a stronger definition that requires that the moment selection procedure is consistent and also is able to determine when there are not enough correct moment conditions to identify c^0.

All limits considered here and below are limits "as $n \to \infty$." Let "$\to_p$" denote "convergence in probability as $n \to \infty$". Let "wp $\to 1$" abbreviate "with probability that goes to one as $n \to \infty$."

We say that a moment selection estimator $\widehat{c} \in \mathcal{C}$ is *consistent* if

$$\widehat{c} = c^0 \text{ wp } \to 1 \text{ under } P^0, \ \forall P^0 \in \mathcal{P} \text{ that satisfy Assumption ID}c^0. \tag{2.9}$$

Because $\mathcal{C}$ is finite, $\widehat{c} = c^0$ wp $\to 1$ is equivalent to the standard (weak) consistency condition that $\hat{c} \to_p c^0$.

We say that a moment selection estimator $\widehat{c} \in \mathcal{C}$ is *s-consistent* if $\widehat{c}$ is consistent and

$$|\widehat{c}| \leq p \text{ wp } \to 1 \text{ under } P^0, \ \forall P^0 \in \mathcal{P} \text{ for which } |c| \leq p \text{ for all } c \in \mathcal{MZ}^0. \tag{2.10}$$

The second part of the definition of s-consistency requires that the moment selection procedure is able to determine whether or not there are one or more over-identifying restrictions, which is necessary for identification of c^0.

The above definitions of consistency and s-consistency are "weak" versions that require behavior that holds "wp$\to$ 1." In Andrews (1997b), we consider *strong* consistency and *strong* s-consistency that require analogous behavior that holds "for n sufficiently large with probability one."

2.6. Performance When Assumption IDc^0 Fails

Below we analyze the behavior of the moment selection procedures introduced below in the case where Assumption IDc^0 does not hold. For this purpose, we make the following definitions. Define

$$\mathcal{CZ}^0 = \mathcal{C} \cap \mathcal{Z}^0. \tag{2.11}$$

As defined, $\mathcal{CZ}^0$ is the set of selection vectors in the parameter space $\mathcal{C}$ that select only moment conditions that equal zero asymptotically for some $\theta \in \Theta$. Define

$$\mathcal{MCZ}^0 = \{c \in \mathcal{CZ}^0 : |c| \geq |c^*| \ \forall c^* \in \mathcal{CZ}^0\}. \tag{2.12}$$

As defined, $\mathcal{MCZ}^0$ is the set of selection vectors in $\mathcal{CZ}^0$ that maximize the number of selected moments out of selection vectors in $\mathcal{CZ}^0$. We show below that for many moment selection procedures discussed below $\widehat{c} \in \mathcal{MCZ}^0$ *wp* $\to 1$ whether or not Assumption IDc^0 holds. That is, for these procedures, with probability that goes to one as $n \to \infty$, $\widehat{c}$ lies in the set of selection vectors that *maximize* the number of selected moments out of all selection vectors in the parameter space $\mathcal{C}$ that select only moments that equal *zero* asymptotically for some $\theta \in \Theta$.

2.7. Basic Assumption

We now state the basic assumption under which the results below hold. This assumption holds quite generally.

Assumption 1. (a) $G_n(\theta) = G^0(\theta) + O_p(n^{-1/2})$ *under* P^0 $\forall \theta \in \Theta \subset R^p$ *for some* R^r*-valued function* $G^0(\cdot)$ *on* Θ, $\forall P^0 \in \mathcal{P}$.
(b) $W_{nc} \to_p W_c^0$ *under* P_0 *for some positive definite matrix* W_c^0 $\forall c \in \mathcal{C}$, $\forall P_0 \in \mathcal{P}$.
(c) $\inf_{\theta\in\Theta} G_{nc}(\theta)' W_{nc} G_{nc}(\theta) \to_p \inf_{\theta\in\Theta} G_c^0(\theta)' W_c^0 G_c^0(\theta) = G_c^0(\theta^*)' W_c^0 G_c^0(\theta^*)$ *under* P^0 *for some* $\theta^* \in \Theta$ *that may depend on* c *and* P^0, $\forall c \in \mathcal{C}$, $\forall P^0 \in \mathcal{P}$.

Assumption 1(a) typically holds by a central limit theorem (CLT) because $G_n(\theta)$ is often a sample average. Assumption 1(b) is a standard condition used to obtain consistency of GMM estimators. It is satisfied by all reasonable choices of weight matrices W_{nc}.

Assumption 1(c) is implied by Assumption 1(b) and the following: $G_n(\theta) \to_p G^0(\theta)$ uniformly over $\theta \in \Theta$ under P^0 for some R^r-valued function $G^0(\cdot)$ that is continuous on Θ, where $\Theta \subset R^p$ is compact, $\forall P^0 \in \mathcal{P}$. The latter can be verified using a generic uniform convergence result, such as a uniform weak LLN, e.g., see Andrews (1992). Alternatively, when the moment conditions are linear in θ, Assumption 1(c) typically holds under almost the same conditions as Assumption 1(a), because the "inf's over $\theta \in \Theta$" can be calculated explicitly. In the linear case, the parameter space Θ can be unbounded.

For illustrative purposes, we provide a sufficient condition for Assumption 1 for the case of stationary data. (The proof of sufficiency is given in the Appendix of Proofs.) Let E^0 denote expectation under P^0. Let $||B||$ denote the Euclidean norm of a vector or matrix, i.e., $||B|| = (tr\ B'B)^{1/2}$.

Assumption STAT. (a) $\{Z_i : i = ..., 0, 1, ...\}$ *is a doubly infinite stationary and ergodic sequence under* P^0, $\forall P^0 \in \mathcal{P}$.
(b) $G_n(\theta) = \frac{1}{n}\sum_{i=1}^n m(Z_i, \theta)$ *and* $m(z, \theta)$ *is continuous in* θ *on* Θ *for all* z *in the support of* Z_i.
(c) $E^0||m(Z_i, \theta)||^2 < \infty$ *and* $\sum_{j=1}^\infty (E^0||E^0(m(Z_i, \theta)|\mathcal{F}_{i-j})||^2)^{1/2} < \infty$ $\forall \theta \in \Theta$, $\forall P^0 \in \mathcal{P}$, *where* $\mathcal{F}_i$ *denotes the* σ*-field generated by* $(..., Z_{i-1}, Z_i)$.
(d) *Either* (i) $\Theta \subset R^p$ *is compact and* $E^0 \sup_{\theta\in\Theta} ||m(Z_i, \theta)|| < \infty$ $\forall P^0 \in \mathcal{P}$ *or* (ii) $m(z, \theta) = m_1(z) + m_2(z)\theta$ $\forall \theta \in \Theta$, *where* $m_1(z) \in R^r$ *and* $m_2(z) \in R^{r\times p}$, *and* $\Theta = R^p$.
(e) Assumption 1(b) holds.

Note that the leading example where the moment conditions are linear in θ and Assumption STAT(d) part (ii) holds is the linear IV estimator of the linear regression model $Y_i = X_i'\theta^* + U_i$ with IV vector $\widetilde{Z}_i \in R^r$. In this case, the moment conditions are $G_n(\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\theta)\widetilde{Z}_i = m_1(Z_i) + m_2(Z_i)\theta$, where $m_1(Z_i) = Y_i\widetilde{Z}_i \in R^r$, $m_2(Z_i) = \widetilde{Z}_i X_i' \in R^{r\times p}$, and $Z_i = (Y_i, X_i', \widetilde{Z'}_i)'$.

3. Moment Selection Criteria

Here we introduce a class of moment selection criteria (MSC) that are analogous to the well-known model selection criteria used for choosing between competing models.

The MSC estimator, $\widehat{c}_{MSC}$, is the value that minimizes $MSC_n(c)$ over $\mathcal{C}$, where

$$MSC_n(c) = J_n(c) - h(|c|)\kappa_n. \tag{3.1}$$

The function $h(\cdot)$ and the constants $\{\kappa_n : n \geq 1\}$ in the definition of $MSC_n(c)$ are specified by the researcher. They are assumed to satisfy:

Assumption MSC. (a) $h(\cdot)$ *is strictly increasing.*
(b) $\kappa_n \to \infty$ *and* $\kappa_n = o(n)$.

Given Assumption MSC, $h(|c|)\kappa_n$ is a "bonus term" that rewards selection vectors that utilize more moment conditions. This term is necessary to offset the increase in $J_n(c)$ that typically occurs when more moment conditions are added even if they are correct moment conditions. Assumption MSC(b) implies that the bonus given for more moment conditions increases without bound as the sample size n increases.

It is always possible to specify moment selection criteria for which Assumption MSC holds, because the researcher chooses $h(\cdot)$ and $\{\kappa_n : n \geq 1\}$.

Now we introduce three examples of moment selection criteria. These are analogues of the BIC, AIC, and HQIC criteria developed for model selection. We refer to them as the GMM–BIC, GMM-AIC, and GMM-HQIC criteria. In each case, they take $h(x) = x - p$. They are defined by

$$\begin{aligned}
\text{GMM} - \text{BIC} \quad &: \quad \kappa_n = \ln n \text{ and } MSC_{\text{BIC},n}(c) = J_n(c) - (|c| - p)\ln n, \\
\text{GMM} - \text{AIC} \quad &: \quad \kappa_n = 2 \text{ and } MSC_{\text{AIC},n}(c) = J_n(c) - 2(|c| - p), \\
\text{GMM} - \text{HQIC} \quad &: \quad \kappa_n = Q \ln\ln n \text{ for some } Q > 2 \text{ and} \\
MSC_{\text{HQIC},n}(c) \quad &= \quad J_n(c) - Q(|c| - p)\ln\ln n.
\end{aligned} \tag{3.2}$$

We show in Section 6 that these are the proper analogues of the BIC, AIC, and HQIC model selection criteria. The GMM–BIC and GMM-HQIC procedures satisfy Assumption MSC. The GMM–AIC procedure does not satisfy Assumption MSC(b) because $\kappa_n = 2 \nrightarrow \infty$. In consequence, the GMM–AIC procedure is not consistent. For brevity, we do not prove this here. The proof is similar to the proof of the lack of consistency of the AIC model selection procedure, see Shibata (1976) and Hannan (1980, 1982). The GMM–AIC procedure has positive probability even asymptotically of selecting too few moments.

Consistency of $\widehat{c}_{MSC}$ is established in the following theorem.

Theorem 1. *Suppose Assumptions* 1 *and* MSC *hold. Then,*
(a) $\widehat{c}_{MSC} \in \mathcal{MCZ}^0$ *wp* $\to 1, \forall P^0 \in \mathcal{P}$,
(b) $\widehat{c}_{MSC} = c^0$ *wp* $\to 1$ *iff Assumption IDc*^{0} *holds and* $c^0 \in \mathcal{C}$, $\forall P^0 \in \mathcal{P}$, *and*
(c) $\widehat{c}_{MSC}$ *is consistent iff* $\widehat{c}_{MSC}$ *is s-consistent iff* $c^0 \in \mathcal{C}$ *for all* $P^0 \in \mathcal{P}$ *for which Assumption IDc*^{0} *holds.*

Comments. 1. Part (a) is a robustness result that specifies the asymptotic behavior of $\widehat{c}_{MSC}$ for all $P^0 \in \mathcal{P}$. Note that if $\mathcal{MCZ}^0 \cap \mathcal{MZ}^0 \neq \emptyset$, then $\widehat{c}_{MSC} \in \mathcal{MZ}^0$ $wp \to 1$, $\forall P^0 \in \mathcal{P}$. The result of part (a) is analogous to results concerning the behavior of extremum estimators when the standard identification condition fails.

2. It is shown in Andrews (1997b) that the results of Theorem 1 hold under somewhat stronger assumptions with "$wp \to 1$" replaced by "for n sufficiently large almost surely." Thus, GMM–BIC and GMM–HQIC are *strongly* consistent and *strongly* s-consistent under suitable assumptions. For GMM-HQIC, these assumptions include the requirement that an asymptotically optimal weight matrix is employed.

3. Theorem 1(b) is similar to Theorem 3 of Hannan (1980) for (weak) consistency of model selection criteria for lag selection in ARMA models.

4. Over-rejection of the J test in finite samples (see the July 1996 issue of the *Journal of Business and Economic Statistics)* affects the MSC only if the amount of over-rejection differs for different selection vectors c. Greater over-rejection for selection vectors with large $|c|$, which seems plausible, leads to a higher probability of using only correct moments, but not necessarily all of them.

5. The proofs of Theorem 1 and other results below are given in the Appendix of Proofs.

4. Downward Testing Procedures

The testing procedures considered in this section and the next are moment selection procedures that formalize the procedures that empirical researchers often use in a less formal, and sometimes vague, fashion. Two advantages of considering precisely specified moment selection procedures are that (i) sufficient conditions for consistency can be established and (ii) the effect of moment selection on post-selection statistical inference can be assessed, e.g., via simulations or the use of the bootstrap.

We consider tests based on the statistic $J_n(c)$. Starting with vectors $c \in \mathcal{C}$ for which $|c|$ is the largest, we carry out tests with progressively smaller $|c|$ until we find a test that does not reject the null hypothesis that the moment conditions considered are all correct. Let $\widehat{k}_{DT}$ be the value of $|c|$ for the first test we find that does not reject. (There is such a *first* test because the J test statistic based on $c = \mathbf{0}$ equals zero.) Given $\widehat{k}_{DT}$, we take the downward testing estimator $\widehat{c}_{DT}$ of c_0 to be the vector that minimizes $J_n(c)$ over $c \in \mathcal{C}$ with $|c| = \widehat{k}_{DT}$. This is the *downward testing* moment selection procedure.

Note that the downward testing moment selection procedure progresses from the most restrictive model to the least restrictive model. This contrasts with a downward testing model selection procedure in which the largest parameter vector, and hence the least restrictive model, is considered first. Upward testing model selection procedures, which are analogous to downward testing moment selection procedures, are referenced in Amemiya (1980) and Pötscher (1989).

We now define $\widehat{k}_{DT}$ and $\widehat{c}_{DT}$ more precisely. Let $\gamma_{n,k} > 0$ denote the critical value employed with the test statistic $J_n(c)$ when $|c| = k$ and the sample size is n. In the leading case where $J_n(c)$ is constructed using an asymptotically optimal weight

matrix, $J_n(c)$ has an asymptotic chi-square distribution with $|c| - \min(p, |c|)$ degrees of freedom when all moment conditions in c are correct.[5] In this case, one can take

$$\gamma_{n,k} = \chi^2_{k-p}(\alpha_n) \tag{4.1}$$

for values of $k > p$, where $\chi^2_{k-p}(\alpha_n)$ denotes $(1-\alpha_n)$-th quantile of a chi-squared distribution with $k-p$ degrees of freedom.

Let $\widehat{k}_{DT} \in [0, r]$ be such that $\min_{c \in \mathcal{C}:|c|=k} J_n(c) > \gamma_{n,k}$ $\forall k = \widehat{k}_{DT}+1, ..., r$ and $\min_{c \in \mathcal{C}:|c|=\widehat{k}_{DT}} J_n(c) \leq \gamma_{n,\widehat{k}_{DT}}$. Define $\widehat{c}_{DT}$ to be any vector in $\mathcal{C}$ for which $|\widehat{c}_{DT}| = \widehat{k}_{DT}$ and $J_n(\widehat{c}_{DT}) = \min_{c \in \mathcal{C}:|c|=\widehat{k}_{DT}} J_n(c)$. In words, $\widehat{k}_{DT}$ is the greatest number of moments for which some $J_n(c)$ test does not reject for some $c \in \mathcal{C}$. Given $\widehat{k}_{DT}$, $\widehat{c}_{DT}$ is the vector that minimizes $J_n(c)$ over vectors $c \in \mathcal{C}$ with $|c| = \widehat{k}_{DT}$.

For consistency of $\widehat{c}_{DT}$, we assume the critical values $\gamma_{n,k}$ satisfy:

Assumption T. *$\gamma_{n,k} \to \infty$ and $\gamma_{n,k} = o(n)$ $\forall k = p+1, ..., r$.*

Assumption T holds if $\{\gamma_{n,k} : k = p+1, ..., r\}$ are defined as in (4.1) with the significance level α_n satisfying $\alpha_n \to 0$ and $\ln \alpha_n = o(n)$ (see Theorem 5.8 of Pötscher (1983)). For example, the latter condition holds if $\alpha_n \geq \lambda_0 \exp(-\lambda_n n)$, for some $0 < \lambda_n \to 0$ and $\lambda_0 > 0$. Note that, when comparing two sets of moment conditions, a consistent MSC of Section 3 implicitly specifies a significance level that also goes to zero as $n \to \infty$ for a test based on the difference between the two $J_n(c)$ statistics.

Consistency of $\widehat{c}_{DT}$ is established in the following theorem.

Theorem 2. *Suppose Assumptions* 1 *and* T *hold. Then,*
(a) $\widehat{c}_{DT} \in \mathcal{MCZ}^0$ wp $\to 1$, $\forall P^0 \in \mathcal{P}$,
(b) $\widehat{c}_{DT} = c^0$ wp $\to 1$ iff Assumption IDc^0 holds and $c^0 \in \mathcal{C}$, $\forall P^0 \in \mathcal{P}$, and
(c) $\widehat{c}_{DT}$ is consistent iff $\widehat{c}_{DT}$ is s-consistent iff $c^0 \in \mathcal{C}$ for all $P^0 \in \mathcal{P}$ for which Assumption IDc^0 holds.

Comments. 1. Theorem 2(b) and Theorem 3(b) below are similar to Theorem 5.7 of Pötscher (1983) for (weak) consistency of upward LM tests for lag selection in ARMA models.

2. Over-rejection by the J test in finite samples, which has been documented in some cases, leads to a higher probability of using only correct moments, but not necessarily all of them.

5. Upward Testing Procedures

Upward testing procedures are based on the statistic $J_n(c)$ and critical values $\{\gamma_{n,k} : k = 1, ..., r\}$. Starting with vectors $c \in \mathcal{C}$ which have the smallest positive values of $|c|$, we carry out tests with progressively larger $|c|$ until we find that all tests with the same value of $|c|$ reject the null hypothesis that the moment conditions considered are all correct. Let $\widehat{k}_{UT}$ denote the largest value such that for all $k \leq \widehat{k}_{UT}$ there is at least one $c \in \mathcal{C}$ with $|c| = k$ for which the null hypothesis is not rejected.

Given $\widehat{k}_{UT}$, we take the upward testing estimator $\widehat{c}_{UT}$ of c_0 to be the vector that minimizes $J_n(c)$ over $c \in \mathcal{C}$ with $|c| = \widehat{k}_{UT}$.

More precisely, $\widehat{k}_{UT}$ and $\widehat{c}_{UT}$ are defined as follows. Let $\mathcal{K} = \{|c| : c \in \mathcal{C}\}$. Define $\widehat{k}_{UT}$ to be the largest integer in $\mathcal{K}$ for which $\min_{c \in \mathcal{C}:|c|=k} J_n(c) \leq \gamma_{n,k}$ $\forall k \in \mathcal{K}$ with $k \leq \widehat{k}_{UT}$. Define $\widehat{c}_{UT}$ to be any vector in $\mathcal{C}$ for which $|\widehat{c}_{UT}| = \widehat{k}_{UT}$ and $J_n(\widehat{c}_{UT}) = \min_{c \in \mathcal{C}:|c|=\widehat{k}_{UT}} J_n(c)$.

As with the downward testing procedure, the critical values $\{\gamma_{n,k} : k = p+1, ..., r\}$ can be taken as in (4.1). For consistency of $\widehat{c}_{UT}$, the critical values are assumed to satisfy Assumption T.

To ensure that the upward testing procedure does not stop at too small a value $|c|$, we need to assume that $\mathcal{C}$ satisfies the following assumption. Let $|\mathcal{MC}^0|$ denote the (unique) number of moments selected by the vectors in $\mathcal{MC}^0$.

Assumption UT. *$\forall k \in \mathcal{K}$ with $k < |\mathcal{MC}^0|$, $\exists c_k \in \mathcal{C}^0$ with $|c_k| = k$, $\forall P^0 \in \mathcal{P}$.*

The parameter space $\mathcal{C}$ can always be defined so that Assumption UT holds, but neither the MSC procedure nor the DT procedure requires this assumption.

Note that it follows from the definitions of $\widehat{c}_{UT}$ and $\widehat{c}_{DT}$ that $|\widehat{c}_{UT}| \leq |\widehat{c}_{DT}|$. Thus, if the UT and DT procedures select different moments, the UT procedure selects fewer than the DT procedure.

Consistency of $\widehat{c}_{UT}$ is established in the following theorem.

Theorem 3. *Suppose Assumptions* 1, T, *and UT hold. Then,*
(a) $\widehat{c}_{DT} \in \mathcal{MCZ}^0$ wp $\to 1$, $\forall P^0 \in \mathcal{P}$,
(b) $\widehat{c}_{DT} = c^0$ wp $\to 1$ iff Assumption IDc^0 holds and $c^0 \in \mathcal{C}$, $\forall P^0 \in \mathcal{P}$, and
(c) $\widehat{c}_{DT}$ is consistent iff $\widehat{c}_{DT}$ is s-consistent iff $c^0 \in \mathcal{C}$ for all $P^0 \in \mathcal{P}$ for which Assumption IDc^0 holds.

6. The Analogy Between BIC/AIC/HQIC and GMM–BIC/GMM–AIC/GMM–HQIC

In this section, we show that GMM–BIC, GMM–AIC, and GMM-HQIC are the proper moment selection analogues of the BIC, AIC, and HQIC model selection procedures.

Consider a log likelihood function $\ell_n(\gamma)$ that depends on a parameter $\gamma \in R^r$. Different models are obtained by setting different elements of γ equal to zero. The maximum likelihood (ML) estimators of γ for different models are just the estimators that maximize the log likelihood function subject to different restrictions on which elements of γ are equal to zero. Let $\widehat{\gamma}_m$ denote the ML estimator of γ for model m, where $m = 1, ..., M$ indexes the models considered. Let q_m denote the number of elements of γ that are set equal to zero in model m. Let $\mathcal{M} = \{1, ..., M\}$ denote the set of models.

The model selection criteria or information criteria (IC) that we consider are of the following form. One chooses the model $m \in \mathcal{M}$ that maximizes

$$IC_n(m) = \ell_n(\widehat{\gamma}_m) - \tfrac{1}{2}(r - q_m)\kappa_n. \tag{6.2}$$

Note that $r - q_m$ equals the number of parameters in model m. The following choices of κ_n yield the BIC, AIC, and HQIC criteria: $\kappa_n = \ln n$ for BIC, $\kappa_n = 2$ for AIC, and $\kappa_n = Q \ln \ln n$ for HQIC, where $Q > 2$.

We consider the asymptotic behavior of $IC_n(m)$ for models that are correct, but not necessarily parsimonious. We aim to elucidate the trade-off that the $IC_n(m)$ procedure makes between the value of the likelihood for correct models and the penalty that it imposes for redundant parameters. Our GMM MSC procedures are designed to provide the same trade-off..

We suppose the likelihood function is regular in the sense that it has a quadratic approximation around the true parameter value γ^0. We suppose m is a correct, but not necessarily parsimonious, model. Then, under standard regularity conditions (see the Appendix of Proofs), we have

$$\ell_n(\widehat{\gamma}_m) = -\tfrac{1}{2}\widetilde{J}_n(m) + S_n, \text{ where } \widetilde{J}_n(m) \xrightarrow{d} \chi^2_{q_m} \tag{6.3}$$

and S_n is a random variable that does not depend on m. See (7.12) for the definitions of S_n and $\widetilde{J}_n(m)$.

Using (6.2) and (6.3), we can write $IC_n(m) = -(\widetilde{J}_n(m) - q_m\kappa_n)/2 + S_n - r\kappa_n/2$. Note that $S_n - r\kappa_n/2$ is a shift random variable that does not depend on m and, hence, has no effect on the outcome of the selection procedure. Thus, maximizing $IC_n(m)$ over models m that are correct, but not necessarily parsimonious, is equivalent to minimizing

$$\widetilde{J}_n(m) - q_m\kappa_n, \tag{6.4}$$

where $\widetilde{J}_n(m)$ is asymptotically $\chi^2_{q_m}$ and q_m is the number of redundant parameters that model m sets equal to zero.[6]

We now turn to the moment selection criteria introduced in Section 3 and show that they are of the same form as the model selection criteria in (6.4). We do this by showing that the choice between different vectors of moment conditions can be reinterpreted as the choice between parameter vectors with different numbers of parameters. This reinterpretation is related to the work of Back and Brown (1993), who address a quite different problem.

The idea of the reinterpretation is as follows. Consider a moment selection vector c. The GMM criterion function for c, viz., $G_{nc}(\theta)'W_{nc}G_{nc}(\theta)$, deletes the moment conditions in $G_n(\theta)$ that correspond to coordinates j for which $c_j = 0$. Alternatively, suppose we retain all moment conditions, but add an unknown mean parameter μ_j to each moment condition $G_{nj}(\theta)$ for which $c_j = 0$. That is, the j-th moment condition is taken to be $G_{nj}(\theta) - \mu_j$ for all j with $c_j = 0$. Then, we treat the parameter vector to be estimated to be the vector that includes θ and the μ_j mean parameters for $j = 1, ..., r - |c|$.

We show below that the minimized value of the GMM criterion function is the same whether one deletes moments or one augments the criterion function with corresponding mean parameters, provided the weight matrices are defined in an asymptotically optimal fashion.[7] Thus, the values of $J_n(c)$ for different vectors c equal the values of a single function that is minimized over parameter vectors with different

numbers of parameters. The latter is analogous to the minimization of the likelihood function over parameter vectors with different numbers of parameters, which is the basis of the BIC, AIC, and HQIC model selection criteria discussed above.

We now state more precisely the result described in the previous paragraph. Define

$$J_n^*(c) = n \inf_{\theta \in \Theta, \mu \in R^{r-p}} (G_n(\theta) - D_c\mu)' W_n (G_n(\theta) - D_c\mu), \tag{6.5}$$

where D_c is an $r \times (r-p)$ duplication matrix such that $[D_c\mu]_j = 0$, $\forall j$ with $c_j = 1$, $[D_c\mu]_j = \mu_1$ for the smallest j with $c_j = 0$, $[D_c\mu]_j = \mu_2$ for the second smallest j with $c_j = 0, ..., [D_c\mu]_j = \mu_{r-|c|}$ for the largest j with $c_j = 0$, and $\mu_{r-|c|+1} = \cdots = \mu_{r-p} = 0$ whenever $|c| > p$.

Note that the number of parameters with selection vector c is $p + r - |c|$, where p is the dimension of θ and $r - |c|$ is the number of excluded moment conditions. We can rewrite the number of parameters with selection vector c as $r - q_c$, where $q_c = |c| - p$. Here q_c is the number of "over-identifying restrictions."

Now, let $\gamma = (\theta', \mu')' \in R^r$. With the selection vector c, the last q_c parameters in γ are set equal to zero in (6.5). Thus, different selection vectors correspond to the setting of different parameters equal to zero in (6.5), just as different models correspond to the setting of different parameters equal to zero in the log likelihood function in (6.2) or (6.3).

Suppose the weight matrices W_{nc} of the GMM criterion function are defined as follows. Let V_n be a consistent estimator of $V = \lim_{n\to\infty} \text{Var}(\sqrt{n} G_n(\theta^0))$. Let V_{nc} denote the $|c| \times |c|$ matrix that equals V_n with the rows and columns of V_n deleted that correspond to elements of c that are zeros. Suppose $W_{nc} = V_{nc}^{-1} + o_p(1)$. Asymptotically optimal weight matrices are of this form. The $o_p(1)$ term allows W_{nc} to be constructed with different estimators of θ^0 for different moment selection vectors c. For $c = \mathbf{0}$, take $W_{nc} = 0$.

We show in the Appendix of Proofs that for W_{nc} as above, we have

$$J_n(c) = J_n^*(c) + o_p(1) \tag{6.6}$$

for all $c \in \mathcal{CZ}^0$. Furthermore, under standard regularity conditions (such as those of Hansen (1982) or Andrews (1997a)), we have: $J_n^*(c) = J_n(c) + o_p(1) \to_d \chi^2_{q_c}$, $\forall c \in \mathcal{CZ}^0$ with $|c| \geq p$. Hence, the moment selection criterion $MSC_n(c)$ of (3.1) with $h(x) = x - p$ satisfies

$$MSC_n(c) = J_n^*(c) - (|c| - p)\kappa_n + o_p(1) = J_n^*(c) - q_c\kappa_n + o_p(1), \tag{6.7}$$

where $J_n^*(c)$ is asymptotically $\chi^2_{q_c}$ for all correct moment selection vectors c and q_c is the number of redundant parameters in γ that the selection vector c sets equal to zero.

The trade-off between the magnitude of the minimized criterion function and the number of parameters is the same in (6.7) as in (6.4). In this sense, the BIC/AIC/HQIC and GMM–BIC/GMM-AIC/GMM-HQIC procedures are analogous.

7. Appendix of Proofs

7.1. Lemma 1 and Its Proof

Lemma 1. *Assumption STAT implies Assumption 1.*

Proof of Lemma 1. Assumption 1(a) holds with $G^0(\theta) = E^0 m(Z_i, \theta)$ by the CLT given in the Theorem of Heyde (1975) (of which there is only one) using Assumption STAT(a), (b), and (c). Assumption 1(c) holds in the case of a nonlinear $m(z, \theta)$ function (i.e., under part (i) of Assumption STAT(d)) by the sufficient condition given following Assumption 1. The latter holds by the ergodic theorem and the uniform weak LLN given in Theorem 4 of Andrews (1992) using Assumption TSE-1D of that paper. In the case of a linear $m(z, \theta)$ function (i.e., under part (ii) of Assumption STAT(d)), Assumption 1(c) holds by computing the infimum over $\theta \in \Theta$ explicitly and applying the ergodic theorem to each component. More specifically, we have

$$\begin{aligned}
&\inf_{\theta \in \Theta} G_{nc}(\theta)' W_{nc} G_{nc}(\theta) = G_{nc}(\widehat{\theta}(c))' W_{nc} G_{nc}(\widehat{\theta}(c)) \\
&\xrightarrow{p} G_c^0(\theta^0(c))' W_c^0 G_c^0(\theta^0(c)) = \inf_{\theta \in \Theta} G_c^0(\theta)' W_c^0 G_c^0(\theta) \text{ under } P^0, \text{ where} \\
&\widehat{\theta}(c) = \left(\tfrac{1}{n} \sum_{i=1}^n m_{2c}(Z_i)' W_{nc} \tfrac{1}{n} \sum_{i=1}^n m_{2c}(Z_i) \right)^{-1} \tfrac{1}{n} \sum_{i=1}^n m_{2c}(Z_i)' W_{nc} \tfrac{1}{n} \sum_{i=1}^n m_{1c}(Z_i) \text{ and} \\
&\theta^0(c) = (E^0 m_{2c}(Z_i)' W_c^0 E^0 m_{2c}(Z_i))^{-1} E^0 m_{2c}(Z_i)' W_c^0 E^0 m_{1c}(Z_i). \quad \square
\end{aligned} \tag{7.1}$$

7.2. Proof of Theorem 1

First, we establish Theorem 1(a). For any $c \in \mathcal{C}$ with $c \notin \mathcal{CZ}^0$, we have

$$J_n(c)/n \xrightarrow{p} \inf_{\theta \in \Theta} G_c^0(\theta)' W_c^0 G_c^0(\theta) > 0 \text{ under } P^0, \tag{7.2}$$

where the convergence holds by Assumption 1(c) and the inequality holds because (i) $G_c^0(\theta) \neq \mathbf{0}$ $\forall \theta \in \Theta$ by the supposition that $c \notin \mathcal{CZ}^0$ and (ii) W_c^0 is positive definite by Assumption 1(b). Equation (7.2) and Assumption MSC(b) yield: For any $c \in \mathcal{C}$ with $c \notin \mathcal{CZ}^0$,

$$MSC_n(c)/n = J_n(c)/n - h(|c|)\kappa_n/n \xrightarrow{p} \inf_{\theta \in \Theta} G_c^0(\theta)' W_c^0 G_c^0(\theta) > 0 \text{ under } P^0. \tag{7.3}$$

For any $c \in \mathcal{CZ}^0$, we have

$$J_n(c) = O_p(1) \text{ under } P^0, \tag{7.4}$$

using Assumptions 1(a) and (c) and the fact that $G_c^0(\theta) = \mathbf{0}$ for some $\theta \in \Theta$. Equation (7.4) and Assumption MSC(b) yield: For any $c \in \mathcal{CZ}^0$,

$$MSC_n(c)/n = O_p(1) - h(|c|)\kappa_n/n = O_p(1) \text{ under } P^0. \tag{7.5}$$

Equations (7.3) and (7.5) imply that $\widehat{c}_{MSC} \in \mathcal{CZ}^0$ $wp \to 1$.

Now, suppose $c_1, c_2 \in \mathcal{CZ}^0$, $c_1 \notin \mathcal{MCZ}^0$, and $c_2 \in \mathcal{MCZ}^0$. Then, $|c_1| < |c_2|$ and by Assumption MSC

$$(h(|c_1|) - h(|c_2|))\kappa_n \to -\infty. \tag{7.6}$$

Equations (7.4) and (7.6) imply that $MSC_n(c_1) > MSC_n(c_2)$ $wp \to 1$. Thus, $\widehat{c}_{MSC} \in \mathcal{MCZ}^0$ $wp \to 1$, as stated in Theorem 1(a).

Now, Assumption IDc^0 and $c^0 \in \mathcal{C}$ imply that $\mathcal{MCZ}^0 = \{c^0\}$. Hence, coupled with Theorem 1(a), the former conditions imply that $\widehat{c}_{MSC} = c^0$ $wp \to 1$. In addition, Assumption IDc^0 is necessary for c^0 to be well defined and $c^0 \in \mathcal{C}$ is necessary for $\widehat{c}_{MSC} = c^0$. Hence, these two conditions are necessary for $\widehat{c}_{MSC} = c^0$ $wp \to 1$ and Theorem 1(b) holds.

To establish Theorem 1(c), we note that by definition s-consistency implies consistency. Consistency implies "$c^0 \in \mathcal{C}$ for all $P^0 \in \mathcal{P}$ for which Assumption IDc^0 holds." In turn, "$c^0 \in \mathcal{C}$ for all $P^0 \in \mathcal{P}$ for which Assumption IDc^0 holds" implies s-consistency because (i) the former implies consistency by Theorem 1(b) and (ii) if $P^0 \in \mathcal{P}$ is such that $|c| \leq p$ for all $c \in \mathcal{MZ}^0$, then $|c| \leq p$ for all $c \in \mathcal{MCZ}^0$ (because $\{\mathbf{0}\} \subseteq \mathcal{CZ}^0 \subseteq \mathcal{Z}^0$) and by Theorem 1(a) $\widehat{c}_{MSC} \in \mathcal{MCZ}^0$ wp $\to 1$, which together imply that $|\widehat{c}_{MSC}| \leq p$ wp $\to 1$. □

7.3. Proof of Theorem 2

First, we establish Theorem 2(a). For any $c \in \mathcal{C}$ with $c \notin \mathcal{CZ}^0$, we have

$$J_n(c)/\gamma_{n,|c|} \xrightarrow{p} \infty \text{ under } P^0 \tag{7.7}$$

by (7.2) and Assumption T. Thus, $\widehat{k}_{DT} \leq |\mathcal{MCZ}^0|$, where $|\mathcal{MCZ}^0|$ denotes the (unique) number of moments selected by the vector(s) in $\mathcal{MCZ}^0$.

For any $c \in \mathcal{CZ}^0$, (7.4) and Assumption T yield

$$J_n(c) < \gamma_{n,|c|} \; wp \to 1 \text{ under } P^0. \tag{7.8}$$

In consequence, $\widehat{k}_{DT} = |\mathcal{MCZ}^0|$ wp$\to$ 1. This result and (7.7) imply that $\widehat{c}_{DT} \in \mathcal{MCZ}^0$ wp$\to$ 1 and, hence, Theorem 2(a) holds.

Now, Theorem 2(b) and (c) follow from Theorem 2(a) by the same argument as used above to show that Theorem 1(b) and (c) follow from Theorem 1(a). □

7.4. Proof of Theorem 3

First, we establish Theorem 3(a). By (7.7), we have $\widehat{k}_{UT} \leq |\mathcal{MCZ}^0|$. By (7.8) and Assumption UT, $\forall k \in \mathcal{K}$ with $k < |\mathcal{MCZ}^0|$, $\exists c_k \in \mathcal{CZ}^0$ with $|c_k| = k$ and $J_n(c_k) < \gamma_{n,|c_k|}$ $wp \to 1$. In consequence, $\widehat{k}_{UT} = |\mathcal{MCZ}^0|$. This result and (7.7) imply that $\widehat{c}_{UT} \in \mathcal{MCZ}^0$ wp$\to$ 1 and, hence, Theorem 3(a) holds.

Now, Theorem 3(b) and (c) follow from Theorem 3(a) by the same argument as used above to show that Theorem 1(b) and (c) follow from Theorem 1(a). □

7.5. Proofs of Results of Section 6

Here we show that (6.3) and (6.6) hold under suitable conditions.

We establish (6.6) under Assumption 1 by first showing that $J_n(c) = J_n^*(c)$ for all $c \in \mathcal{C}\mathcal{Z}^0$ when the $o_p(1)$ term appended to the definition of V_{nc} is zero. Given $c \in \mathcal{C}$, we assume without loss of generality that c deletes the last $r - |c|$ moments. Then, we can write

$$
\begin{aligned}
&G_n(\theta) = \begin{pmatrix} G_{nc}(\theta) \\ G_{nc}^*(\theta) \end{pmatrix} \text{ for } G_{nc}(\theta) \in R^{|c|} \text{ and } G_{nc}^*(\theta) \in R^{r-|c|}, \\
&D_c\mu = \begin{pmatrix} \mathbf{0} \\ \mu_* \end{pmatrix} \text{ for } \mathbf{0} \in R^{|c|} \text{ and } \mu_* = (\mu_1, ..., \mu_{r-|c|})' \in R^{r-|c|}, \\
&V_n = \begin{bmatrix} V_{nc} & B \\ B' & D \end{bmatrix} \text{ for } V_{nc} \in R^{|c|\times|c|},\ B \in R^{|c|\times(r-|c|)}, \text{ and } D \in R^{(r-|c|)\times(r-|c|)}, \\
&W_{nc} = V_{nc}^{-1}, \text{ and } W_n = V_n^{-1}. \qquad (7.9)
\end{aligned}
$$

Using the standard formula for the inverse of a partitioned matrix, we obtain

$$
\begin{aligned}
J_n^*(c) &= n \inf_{\theta\in\Theta, \mu_* \in R^{r-|c|}} \begin{pmatrix} G_{nc}(\theta) \\ G_{nc}^*(\theta) - \mu_* \end{pmatrix}' \begin{pmatrix} V_{nc} & B \\ B' & D \end{pmatrix}^{-1} \begin{pmatrix} G_{nc}(\theta) \\ G_{nc}^*(\theta) - \mu_* \end{pmatrix} \\
&= n \inf_{\theta\in\Theta, \mu_* \in R^{r-|c|}} (Q_1(\theta) + Q_2(\theta, \mu_*)), \text{ where } Q_1(\theta) = G_{nc}(\theta)' V_{nc}^{-1} G_{nc}(\theta), \\
Q_2(\theta, \mu_*) &= \begin{pmatrix} G_{nc}(\theta) \\ G_{nc}^*(\theta) - \mu_* \end{pmatrix}' \begin{pmatrix} FEF' & -FE \\ -EF' & E \end{pmatrix} \begin{pmatrix} G_{nc}(\theta) \\ G_{nc}^*(\theta) - \mu_* \end{pmatrix}, \\
&F = V_{nc}^{-1} B, \text{ and } E = (D - B' V_{nc}^{-1} B)^{-1}. \qquad (7.10)
\end{aligned}
$$

Now, we solve for the value $\widehat{\mu}_*(\theta)$ that minimizes $Q_2(\theta, \mu_*)$ for given θ. The solution to this simple quadratic minimization problem is $\widehat{\mu}_*(\theta) = G_{nc}^*(\theta) - F' G_{nc}(\theta)$. Substituting this into (7.10) and simplifying yields $Q_2(\theta, \widehat{\mu}_*(\theta)) = 0$ and $J_n^*(c) = J_n(c)$, which establishes (6.6) when the $o_p(1)$ term added to V_{nc} is zero.

Now, (6.6) with the $o_p(1)$ term present follows from (6.6) without the $o_p(1)$ term provided $\sqrt{n} G_{nc}(\widehat{\theta}_n(c)) = O_p(1)$, because the $o_p(1)$ term adds at most an $o_p(1)$ term to $J_n(c)$ in this case. The previous condition holds under Assumption 1 $\forall c \in \mathcal{Z}^0$.

Next, we show that (6.3) holds under standard ML regularity conditions. Let m denote a correct, but not necessarily parsimonious, model. We partition γ, γ^0, and $\widehat{\gamma}_m$ as $\gamma = (\alpha', \beta')'$, $\gamma^0 = (\alpha^{0\prime}, \beta^{0\prime})'$, and $\widehat{\gamma}_m = (\widehat{\alpha}_m', \mathbf{0}')'$,where $\alpha, \alpha^0, \widehat{\alpha}_m \in R^{r-q_m}$ and $\beta, \beta^0 \in R^{q_m}$. The assumption that m is a correct model implies that $\beta^0 = \mathbf{0}$. The ML estimator $\widehat{\gamma}_m$ for model m sets $\beta = \mathbf{0}$ and maximizes $\ell_n(\alpha) = \ell_n((\alpha', \mathbf{0}')')$ over a parameter space $A \subset R^{r-q_m}$.

We assume the likelihood function is sufficiently regular that the following conditions hold: (i) m is a correct model and $\widehat{\gamma}_m \to_p \gamma^0$, (ii) α^0 is an interior point of A, (iii) $\ell_n(\gamma)$ is twice continuously differentiable at γ^0 with probability one, (iv) $(\partial/\partial\gamma)\ell_n(\gamma^0) \,/\sqrt{n} \to_d N(\mathbf{0}, \mathcal{I})$, where $\mathcal{I}$ is a positive definite $r \times r$ matrix, (v) for some function $\mathcal{I}(\gamma)$ and some $\varepsilon > 0$, $\sup_{\gamma \in B(\gamma^0, \varepsilon)} \left\| -(\partial^2/\partial\gamma\partial\gamma')\ell_n(\gamma) - \mathcal{I}(\gamma) \right\| = o_p(1)$, $\mathcal{I}(\gamma)$ is continuous at γ^0, and $\mathcal{I}(\gamma^0) = \mathcal{I}$, where $B(\gamma^0, \varepsilon)$ is a ball in R^r of radius ε

centered at γ^0. These conditions are sufficiently general to cover many econometric models. The condition that $\ell_n(\gamma)$ is pointwise twice differentiable could be relaxed. For brevity, we do not do so.

We partition the information matrix $\mathcal{I}$ conformably with α and β with diagonal blocks $\mathcal{I}_\alpha$ and $\mathcal{I}_\beta$ and off-diagonal blocks $\mathcal{I}_{\alpha\beta}$ and $\mathcal{I}_{\beta\alpha}$. We obtain

$$\sqrt{n}(\widehat{\alpha}_m - \alpha^0) = \mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\alpha^0) + o_p(1) \tag{7.11}$$

using the first order conditions $(\partial/\partial\alpha)\ell_n(\widehat{\alpha}_m) = \mathbf{0}$ wp $\to 1$, element-by-element mean-value expansions of $(\partial/\partial\alpha)\ell_n(\widehat{\alpha}_m)$ about α^0, and some rearrangements that utilize conditions (iv) and (v).

Now, a two-term Taylor expansion of $\ell_n(\widehat{\gamma}_m)$ about γ^0 gives

$$\begin{aligned}
\ell_n(\widehat{\gamma}_m) &= \ell_n(\gamma^0) + (Z_n^*)'\mathcal{I}_n^*\sqrt{n}(\widehat{\gamma}_m - \gamma^0) - \tfrac{1}{2}\sqrt{n}(\widehat{\gamma}_m - \gamma^0)'\mathcal{I}_n^*\sqrt{n}(\widehat{\gamma}_m - \gamma^0) \\
&= \ell_n(\gamma^0) + \tfrac{1}{2}(Z_n^*)'\mathcal{I}_n^* Z_n^* - \tfrac{1}{2}(\sqrt{n}(\widehat{\gamma}_m - \gamma^0) - Z_n^*)'\mathcal{I}_n^*(\sqrt{n}(\widehat{\gamma}_m - \gamma^0) - Z_n^*) \\
&= S_n - \tfrac{1}{2}\widetilde{J}_n(m), \text{where } S_n = \ell_n(\gamma^0) + \frac{1}{2} Z_n' \mathcal{I} Z_n,\ Z_n = \mathcal{I}^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \gamma} \ell_n(\gamma^0), \\
\widetilde{J}_n(m) &= (\sqrt{n}(\widehat{\gamma}_m - \gamma^0) - Z_n^*)'\mathcal{I}_n^*(\sqrt{n}(\widehat{\gamma}_m - \gamma^0) - Z_n^*) \\
&\quad + Z_n'\mathcal{I} Z_n - (Z_n^*)'\mathcal{I}_n^* Z_n^*,
\end{aligned} \tag{7.12}$$

γ^* is a point on the line segment joining $\widehat{\gamma}_m$ and γ^0, $\mathcal{I}_n^* = -(\partial^2/\partial\gamma\partial\gamma')\ell_n(\gamma^*)/n$, and $Z_n^* = (\mathcal{I}_n^*)^{-1}(\partial/\partial\alpha)\ell_n(\gamma^0)/\sqrt{n}$.

By conditions (i), (iv), and (v), $Z_n'\mathcal{I} Z_n - (Z_n^*)'\mathcal{I}_n^* Z_n^* = o_p(1)$. This result, (7.11), (7.12), and conditions (iv) and (v) give

$$\begin{aligned}
\widetilde{J}_n(m) &= \left(\begin{pmatrix} \mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) \\ \mathbf{0} \end{pmatrix} - Z_n \right)' \mathcal{I} \left(\begin{pmatrix} \mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) \\ \mathbf{0} \end{pmatrix} - Z_n \right) + o_p(1) \\
&= \left(\mathcal{I} \begin{pmatrix} \mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) \\ \mathbf{0} \end{pmatrix} - \frac{1}{\sqrt{n}} \frac{\partial}{\partial \gamma} \ell_n(\gamma^0) \right)' \mathcal{I}^{-1} \left(\mathcal{I} \begin{pmatrix} \mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) \\ \mathbf{0} \end{pmatrix} - \frac{1}{\sqrt{n}} \frac{\partial}{\partial \gamma} \ell_n(\gamma^0) \right) + o_p(1) \\
&= \left(\begin{matrix} \mathbf{0} \\ \mathcal{I}_{\beta\alpha}\mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) - \frac{1}{\sqrt{n}} \frac{\partial}{\partial \beta} \ell_n(\gamma^0) \end{matrix} \right)' \mathcal{I}^{-1} \left(\begin{matrix} \mathbf{0} \\ \mathcal{I}_{\beta\alpha}\mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) - \frac{1}{\sqrt{n}} \frac{\partial}{\partial \beta} \ell_n(\gamma^0) \end{matrix} \right) + o_p(1) \\
&= \left(\mathcal{I}_{\beta\alpha}\mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) - \frac{1}{\sqrt{n}} \frac{\partial}{\partial \beta} \ell_n(\gamma^0) \right)' \left(\mathcal{I}_\beta - \mathcal{I}_{\beta\alpha}\mathcal{I}_\alpha^{-1}\mathcal{I}_{\alpha\beta} \right)^{-1} \\
&\quad \times \left(\mathcal{I}_{\beta\alpha}\mathcal{I}_\alpha^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \alpha} \ell_n(\gamma^0) - \frac{1}{\sqrt{n}} \frac{\partial}{\partial \beta} \ell_n(\gamma^0) \right) + o_p(1) \\
&\xrightarrow{d} \chi^2_{q_m}.
\end{aligned} \tag{7.13}$$

Equations (7.12) and (7.13) establish (6.3).

8. Footnotes

[1]The author thanks four anonymous referees and the coeditor Alain Monfort for very helpful comments and suggestions, especially in terms of the formulation of Assumption IDc^0. The author also thanks Moshe Buchinsky, Peter Hall, Lars Hansen, Ariel Pakes, Walter Philipp, Chris Sims, and Yuichi Kitamura for helpful comments and Glena Ames for typing the manuscript. The author gratefully acknowledges the research support of the National Science Foundation via grant number SBR–9410675.

[2]More generally, for consistency and asymptotic normality, one can take $\widehat{\theta}_n$ to be any value in Θ that yields a value of $G_n(\theta)'W_nG_n(\theta)$ that is within $o_p(n^{-1})$ of the minimum, see Pakes and Pollard (1989).

[3] In this case, W_{nc} is the inverse of an estimator, V_{nc}, of the asymptotic covariance matrix, V_c, of the moment conditions $\sqrt{n}G_{nc}(\theta^0)$. We recommend that V_{nc} be defined using the same general formula for each selection vector c (to minimize the differences across vectors c) and with the sample average of the moment conditions subtracted off. For example, in an iid case with $G_n(\theta) = \frac{1}{n}\sum_{i=1}^n m(Z_i,\theta)$ and $V_c = \text{Var}(m_c(Z_i,\theta^0))$, we recommend defining V_{nc} as follows:

$$V_{nc} = \frac{1}{n}\sum_{i=1}^{n}\left(m_c(Z_i,\widehat{\theta}(c)) - \overline{m}_{nc}(\widehat{\theta}(c))\right)\left(m_c(Z_i,\widehat{\theta}(c)) - \overline{m}_{nc}(\widehat{\theta}(c))\right)',$$

where $\overline{m}_{nc}(\theta) = \frac{1}{n}\sum_{i=1}^n m_c(Z_i,\theta)$ and $\widehat{\theta}(c)$ is some estimator of θ^0. In the case of temporal dependence, sample averages can be subtracted off from a heteroskedasticity and autocorrelation consistent covariance matrix estimator in an analogous fashion. Subtracting off the sample averages is particularly important when some of the moment conditions are not correct.

[4]An arbitrary $o_p(1)$ term can be added to the right-hand side of this equation and (2.6) and (2.7) without affecting any of the results below. This indicates that the infimum over Θ need not be computed exactly.

[5]For conditions under which this result holds, see Hansen (1982) for the case of moment conditions that are smooth in θ and Andrews (1997) for the case of moment conditions that may be non-differentiable and/or discontinuous. Andrews' results extend those of Pakes and Pollard (1989), who do not discuss the $J_n(c)$ statistic.

[6]Of course, when carrying out a model selection procedure, one maximizes $IC_n(m)$ over all models m in $\mathcal{M}$, not just the correct models, because one does not know which models are correct. Here we focus only on the correct models, because we are interested in the relative magnitudes of the maximized log likelihood and the penalty term for redundant parameters for correct models.

[7]Furthermore, by the results of Back and Brown (1993), the GMM estimator $\widehat{\theta}_n(c)$ defined in (2.7) is the same as the estimator of θ that is attained via minimizing the GMM criterion function augmented with corresponding μ_j parameters, provided the weight matrices are defined in an asymptotically optimal fashion.

9. References

Akaike, H. (1969): "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21, 243–247.

_____ (1977): "On Entropy Maximization Principle," in *Applications of Statistics*, ed. by P. R. Krishnaiah. Amsterdam: North–Holland.

Amemiya, T. (1980): "Selection of Regressors," *International Economic Review*, 21, 331-354.

Andrews, D. W. K. (1988): "Laws of Large Numbers for Dependent Non-identically Distributed Random Variables," *Econometric Theory*, 4, 458–467.

_____ (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241–257.

_____ (1997a): "A Stopping Rule for the Computation of Generalized Method of Moments Estimators," *Econometrica*, 65, 913–931.

_____ (1997b): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation: Strong Consistency and Simulation Results," discussion paper, Department of Economics, Yale University.

Eichenbaum, M. S., L. P. Hansen, and K. J. Singleton (1988): "A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice under Uncertainty," *Quarterly Journal of Economics*, 103, 51–78.

Gallant, A. R., D. Hsieh, and G. Tauchen (1997): "Estimation of Stochastic Volatility Models with Diagnostics," *Journal of Econometrics*, forthcoming.

Gallant, A. R. and G. Tauchen (1996): "Which Moments to Match?" *Econometric Theory*, 12, 657–681.

Hannan E. J. (1980): "The Estimation of the Order of an ARMA Process," *Annals of Statistics*, 8, 1071–1081.

_____ (1982): "Testing for Autocorrelation and Akaike's Criterion," in *Essays in Statistical Science*, ed. by J. M. Gani and E. J. Hannan. Sheffield: Applied Probability Trust, pp. 403–412.

Hannan, E. J. and M. Deistler (1988): *The Statistical Theory of Linear Systems*. New York: Wiley.

Hannan, E. J. and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society,* Series B, 41, 190–195.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Heyde, C. C. (1975): "On the Central Limit Theorem and Iterated Logarithm Law for Stationary Processes," *Bulletin of the Australian Mathematical Society*, 12, 1–8.

Kolaczyk, E. D. (1995): "An Information Criterion for Empirical Likelihood with General Estimating Equations," unpublished manuscript, Department of Statistics, University of Chicago.

Lu, B. and D. W. K. Andrews (1997): "Consistent Model and Moment Selection Procedures for GMM Estimation with Applications to Dynamic Panel Models," manuscript under preparation, Cowles Foundation for Research in Economics, Yale University.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

Pesaran, M. H. and R. J. Smith (1994): "A Generalized R^2 Criterion for Regression Models Estimated by the Instrumental Variables Method," *Econometrica*, 62, 705–710.

Pötscher, B. M. (1983): "Order Estimation in ARMA-models by Lagrangian Multiplier Tests," *Annals of Statistics*, 11, 872–885.

_____ (1989): "Model Selection under Nonstationarity: Autoregressive Models and Stochastic Linear Regression Models," *Annals of Statistics*, 17, 1257-1274.

Rissanen, J. (1978): "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Shibata, R. (1976): "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117–126.

Smith, R. J. (1992): "Non-nested Tests for Competing Models Estimated by Generalized Method of Moments," *Econometrica*, 60, 973–980.

Stock, J. H. and J. Wright (1997): "GMM Weak Identification," unpublished manuscript, Kennedy School of Government, Harvard University.

White, H. (1982): "Instrumental Variables Regression with Independent Observations," *Econometrica*, 50, 483–499.