CONSISTENT MODEL AND MOMENT SELECTION PROCEDURES FOR GMM ESTIMATION WITH APPLICATION TO DYNAMIC PANEL DATA MODELS

BY

DONALD W. K. ANDREWS and BIAO LU

COWLES FOUNDATION PAPER NO. 1015

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

2001

http://cowles.econ.yale.edu/

ELSEVIER
Journal of Econometrics 101 (2001) 123–164
JOURNAL OF Econometrics
www.elsevier.nl/locate/econbase

Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models

Donald W.K. Andrews[a,*], Biao Lu[b,1]

[a]*Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven 06520-8281, USA*
[b]*Department of Finance, School of Business Administration, University of Michigan, Ann Arbor, MI 48109, USA*

Received 8 January 1999; received in revised form 12 July 2000; accepted 1 August 2000

Abstract

This paper develops consistent model and moment selection criteria for GMM estimation. The criteria select the correct model specification and all correct moment conditions asymptotically. The selection criteria resemble the widely used likelihood-based selection criteria BIC, HQIC, and AIC. (The latter is not consistent.) The GMM selection criteria are based on the J statistic for testing over-identifying restrictions. Bonus terms reward the use of fewer parameters for a given number of moment conditions and the use of more moment conditions for a given number of parameters. The paper also considers a consistent downward testing procedure. The paper applies the model and moment selection criteria to dynamic panel data models with unobserved individual effects. The paper shows how to apply the selection criteria to select the lag length for lagged dependent variables, to detect the number and locations of structural breaks, to determine the exogeneity of regressors, and/or to determine the existence of correlation between some regressors and the individual effect. To illustrate the finite sample performance of the selection criteria and the testing procedures and their impact

* Corresponding author. Fax: + 1-203-432-6167.

E-mail address: donald.andrews@yale.edu (D.W.K. Andrews).

[1] The first version of this paper was done when the second author was studying at Department of Economics, Yale University. We thank two referees, Peter Phillips, Chris Sims, and seminar participants at Yale University and SUNY at Stony Brook for very helpful comments and suggestions. The first author gratefully acknowledges the research support of the National Science Foundation via grant numbers SBR-9410675 and SBR-9730277.

PII: S0304-4076(00)00077-4

on parameter estimation, the paper reports the results of a Monte Carlo experiment on a dynamic panel data model.

JEL classification: C12; C13; C52

Keywords: Akaike information criterion; Bayesian information criterion; Consistent selection procedure; Generalized method of moments estimator; Instrumental variables estimator; Model selection; Moment selection; Panel data model; Test of over-identifying restrictions

1. Introduction

Many econometric models are specified through moment conditions rather than complete distributional assumptions. Examples are dynamic panel data models with unobserved individual effects and macroeconomic models with rational expectations. Such models are usually estimated using generalized method of moments (GMM), see Hansen (1982). For consistency and asymptotic normality, this method relies on the correct specification of the model and, given the model, on the specification of correct moment conditions. To date, no procedures are available in the literature that consider the problem of selecting the correct model and correct moment conditions in a GMM context.

In this paper, we introduce consistent model and moment selection criteria (MMSC) and downward testing procedures that are able to select the correct model and moments for GMM estimation with probability that goes to one as the sample size goes to infinity. Our results apply to both nested and non-nested models. Our results extend those of Andrews (1999), who considers the problem of selection of correct moments given the correct model. Our results extend the model selection literature, which considers model selection based on the likelihood under full distributional assumptions, to GMM contexts. Our results provide a model selection alternative to the non-nested tests for GMM models considered in Smith (1992).

In the paper, we apply the MMSC and downward testing procedures to dynamic panel data models. We show that these procedures can be used to consistently select from a number of different specifications of the model and moment conditions. The MMSC and testing procedures can be applied to questions of lag length, existence of structural breaks, exogeneity of regressors, and correlation between regressors and an unobserved individual effect. Of course, in any one application, one would not want to try to use the data to answer all of these questions simultaneously. To do so would result in very poor finite sample behavior. Nevertheless, for theoretical purposes, we set up a general model that incorporates all these questions and allows us to provide one set of results that simultaneously covers the many restricted sub-models of interest.

We explore the finite sample properties of the MMSC and testing procedures and their impact on parameter estimation via a Monte Carlo experiment based on a restricted version of the general dynamic panel data model. In this model, the true lag length of the lagged dependent variables is unknown. Furthermore, it is not known whether a regressor is predetermined or strictly exogenous with respect to the time-varying error component or whether the regressor is correlated with the unobserved individual effect.

The MMSC that we consider resemble the widely used BIC, AIC, and HQIC model selection criteria. (See Hannan and Quinn (1979) for the latter.) The MMSC are based on the J test statistic for testing over-identifying restrictions. They include bonus terms that reward the use of more moment conditions for a given number of parameters and the use of less parameters for a given number of moment conditions. The J statistic is an analogue of (minus) the log-likelihood function and the bonus terms are analogues of (minus) the term that penalizes the use of more parameters in a standard model selection criterion.

For illustration, we define the MMSC–BIC here. Setting different elements of θ equal to zero yields different models. For example, in a model with lagged dependent variables, setting different lag coefficients to zero yields models with different numbers of lags. As a second example, suppose one has two competing non-nested models with two corresponding parameter vectors and two sets of GMM estimating equations. Then, the two parameter vectors can be stacked to yield a single parameter θ. Setting the second parameter vector equal to zero yields the first model and vice versa.

Next, let (b, c) denote a pair of model and moment selection vectors. That is, b is a vector that selects some parameters from the vector θ, but not necessarily all of them. And c selects some moments, but not necessarily all of them. Let $|b|$ and $|c|$ denote the numbers of parameters and moments, respectively, selected by (b, c). Let $J_n(b, c)$ denote the J test statistic for testing over-identifying restrictions, constructed using the parameters selected by b and the moments selected by c. Let $\mathscr{BC}$ be the parameter space for the model and moment selection vectors (b, c). Let n denote the sample size. Then, the MMSC–BIC criterion selects the pair of vectors (b, c) in $\mathscr{BC}$ that minimizes

$$J_n(b, c) - (|c| - |b|)\ln n. \tag{1.1}$$

In Andrews and Lu (1999), we show that this criterion is the proper analogue of the BIC model selection criterion in the sense that it makes the same asymptotic trade-off between the ‘model fit’ and the ‘number of parameters’.

The downward testing procedure considered here selects models and moments by carrying out J tests of over-identifying restrictions. The downward testing procedure starts with the model/moment combinations with the most number of over-identifying restrictions and tests the null hypothesis that all moments under test have mean zero for some parameter value. The procedure tests model/moment combinations with progressively fewer over-identifying restrictions until it finds one that does not reject the null hypothesis. This one is the selected model/moment combination.

We now discuss the general dynamic panel data model considered in the paper. The model does not assume specific distributions for the errors in the model. Instead, following many papers in the recent literature, the model is specified by a sequence of assumptions on the means and covariances of the random variables that enter the model. These assumptions imply a sequence of moment conditions that may be used for GMM estimation of the parameters.

The general dynamic panel data model that we consider nests as special cases the models in Hausman and Taylor (1981), Anderson and Hsiao (1982), Bhargava and Sargan (1983), Breusch et al. (1989), Arellano and Bover (1995), and Ahn and Schmidt (1995). In addition, the model shares a common feature with those in Chamberlain (1984) and Holtz-Eakin et al. (1988) in the sense that coefficients can vary over time. The model also incorporates some novel features by allowing for (i) potentially unknown lag length for the lagged dependent variables, (ii) possible structural breaks in the parameters at unknown times, (iii) regressors whose predetermined/strictly exogenous status is unknown, and (iv) regressors whose correlation with the individual effect is not known to be zero or nonzero.

To evaluate the finite sample properties of the MMSC and testing procedures, we conduct a Monte Carlo experiment on a dynamic panel data model that is a restricted version of the general model. The consistent MMSC are shown to have good performance in selecting the correct parameter vector and correct moment conditions. Conducting model and moment selection has an impact on parameter estimation. The post-selection GMM estimators can have much lower biases, standard errors, and root mean squared-errors and more accurate rejection rates than a standard GMM estimator without model and moment selection. We find that the MMSC–BIC and downward testing procedures are the best procedures in all cases considered except that with the smallest sample size.

We now review the literature related to this paper. In addition to Andrews (1999), the closest literature to the model and moment selection results of this paper is that concerning likelihood-based model selection criteria. The AIC criterion was introduced by Akaike (1969). The BIC criterion was introduced by Schwarz (1978), Rissanen (1978), and Akaike (1977). The HQIC criterion was introduced by Hannan and Quinn (1979). The PIC criterion was introduced by Phillips and Ploberger (1996). Consistency, strong consistency, or lack thereof of these procedures are established by Shibata (1976), Hannan (1980, 1982), and Hannan and Deistler (1988), as well as some of the references above. The use of model selection procedures in general non-linear models has been considered by Kohn (1983), Nishii (1988), and Sin and White (1996). The effect of model selection on post-model selection inference is considered by Pötscher (1991), Pötscher and Novák (1994), and Kabaila (1995) among others. For the literature on regressor selection, see Amemiya (1980), Pötscher (1989), and references therein.

Other literature related to this paper includes Kolaczyk (1995), who considers an analogue of the AIC model selection criterion in an empirical likelihood context, and Pesaran and Smith (1994), who consider an R^2-type criterion that can be used for model selection in linear regression models estimated by instrumental variables.

In addition, the results of this paper are related to the test of Eichenbaum et al. (1988) of whether a given subset of moment conditions is correct or not. They propose a likelihood-ratio like test based on the GMM criterion function. The results of this paper also are related to the literature on non-nested tests in GMM contexts, see Smith (1992).

Gallant and Tauchen (1996) 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 et al. (1997) consider using t-ratios for individual moment conditions as diagnostics for moment failure.

Our results for dynamic panel data models follow a long line of research in econometrics. Early contributions including Mundlak (1961), Balestra and Nerlove (1966), and Maddala (1971). More recently, static panel data models with unobserved individual effects that may be correlated with some of the explanatory variables are studied in Hausman and Taylor (1981), Amemiya and MaCurdy (1986), Breusch et al. (1989), and Keane and Runkle (1992). Dynamic panel data models with unobserved individual effects are studied in Anderson and Hsiao (1982), Bhargava and Sargan (1983), Chamberlain (1984), Holtz-Eakin et al. (1988), Arellano and Bond (1991), Ahn and Schmidt (1995), Blundell and Bond (1995), and Arellano and Bover (1995). The latter paper provides a nice summary of many of the models that have been considered in the literature.

The rest of the paper is organized as follows: Section 2 introduces the general model and moment selection problem and defines the ‘correct’ model and moment selection vectors. Section 3 introduces a class of model and moment selection criteria and provides conditions for consistency of these criteria in a general GMM context. Section 4 introduces the downward testing procedure and provides conditions for consistency of this testing procedure. Section 5 specifies a general dynamic model for panel data and compares it to models in the literature. Section 5 also provides an array of different restrictions on the general panel data model, specifies the moment conditions implied by these restrictions, and applies the model and moment selection procedures of Sections 3 and 4 to this model. Section 6 evaluates the finite sample performance of the model and moment selection procedures via Monte Carlo simulation. In this section, a restricted version of the general dynamic panel data model of Section 5 is used. Section 7 concludes. An Appendix contains proofs.

2. The model and moment selection problem

2.1. Introduction

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

We have a random vector of empirical moments

$$G_n(\theta): \Theta \to R^r \qquad (2.1)$$

and a random $r \times r$ weight matrix W_n, both of which depend on $\{Z_i\colon i \leqslant n\}$. The parameter space Θ is a subset of R^p. Typically, the empirical moments are of the form $G_n(\theta) = (1/n)\sum_{i=1}^{n} m(Z_i, \theta)$.

We assume that $G_n(\theta)$ converges in probability as $n \to \infty$ to a function $G^0(\theta)$ $\forall \theta \in \Theta$, $\forall P^0 \in \mathscr{P}$. (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 the entire parameter vector θ is to be estimated and that all r moment conditions are correct. By the latter, we mean that for some $\theta^0 \in \Theta$, one has $G^0(\theta^0) = \mathbf{0}$. 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 $\hat{\theta}_n$ of θ^0 is defined to minimize

$$G_n(\theta)' W_n G_n(\theta) \quad \text{over } \theta \in \Theta. \qquad (2.2)$$

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

Here, we consider the case where the parameter vector θ may incorporate several models. By setting different elements of θ equal to zero, one obtains different models. Two examples of this are given in the Introduction. As a third example, consider a model that may have structural breaks in the parameters (perhaps at some unknown time(s)). The vector θ can include the pre-break values of the parameter plus post-break deviations from the pre-break values. Different sets of post-break deviations can denote changes at different times. If the post-break deviations are set equal to zero, then one obtains the model with no structural breaks.

We consider the case where not all of the moments in $G_n(\theta)$ are necessarily correct. That is, it may be the case that there is no vector $\theta^0 \in \Theta$ for which $G^0(\theta^0) = \mathbf{0}$. This situation can arise for a variety of reasons. It clearly arises in the example of selecting between two non-nested GMM models mentioned in the Introduction. In this case, $G_n(\theta)$ consists of the moments for the two models stacked one on top of the other. In this case, one expects a priori that one set of moments or the other is correct, but not both. Of course, this example extends to the case of more than two non-nested models.

In addition, one may have some incorrect moments when $G_n(\theta)$ consists of moments for a single model or nested models, but there is a hierarchy of restrictions on the model(s). In such cases, some moment conditions may hold, whereas others may not. For example, some moment conditions might hold if certain variables are predetermined and an additional set may hold if, in addition, the variables are strictly exogenous.

By allowing for incorrect moment conditions, as in Andrews (1999), we provide a method of dealing with the common problem in empirical applications that the J test of over-identifying restrictions rejects the null hypothesis that all moment conditions are correct.

Below we show that under certain assumptions it is possible to consistently estimate the 'correct' model and the 'correct' moment conditions given suitable definitions of 'correct'. This allows one to construct a GMM estimator that relies only on the correct model and moment conditions asymptotically, provided there are some over-identifying restrictions on the correct model.

2.2. *Definition of the correct model and moment selection vectors*

Let $(b, c) \in R^p \times R^r$ denote a pair of *model* and *moment selection vectors*. By definition, b and c are each vectors of zeros and ones. If the jth element of b is a one, then the jth element of the parameter vector θ is a parameter to be estimated. If the jth element is a zero, then the jth element of θ is set equal to zero and is not estimated. If the jth element of c is a one, then the jth moment condition is included in the GMM criterion function, whereas if the jth element is a zero, it is not included. Let

$$\begin{aligned}\mathscr{S} = \{&(b, c) \in R^p \times R^r\colon b_j = 0 \text{ or } 1\ \forall 1 \leqslant j \leqslant p,\ c_k = 0 \text{ or } 1\ \forall 1 \leqslant k \leqslant r,\\ &\text{where } b = (b_1, \ldots, b_p)' \text{ and } c = (c_1, \ldots, c_r)'\}. \qquad (2.3)\end{aligned}$$

Let $|b|$ denote the number of parameters to be estimated given b, i.e., $|b| = \sum_{j=1}^{p} b_j$. Let $|c|$ denote the number of moments selected by c, i.e., $|c| = \sum_{k=1}^{r} c_k$.

Consider any p-vector θ, any r-vector v, and any $(b, c) \in \mathscr{S}$ with $c \neq \mathbf{0}$. Let $\theta_{[b]}$ denote the p-vector that results from setting all elements of θ equal to zero whose coordinates equal coordinates of elements of b that are zeros (i.e., $\theta_{[b]}$ is the element by element (Hadamard) product of θ and b). 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 moments that are specified by c. In sum, the subscript $[b]$ sets some elements of a vector equal to zero, whereas the subscript c deletes some elements. For $c = \mathbf{0}$, let $v_c = 0$ $(\in R)$.

We now define the 'correct' model selection vector b^0 and the 'correct' moment selection vector c^0. Let $c^0(\theta)$ be the r vector of zeros and ones whose jth element is one if the jth 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

$$\mathscr{Z}^0 = \{(b,\ c) \in \mathscr{S}\colon\ c = c^0(\theta) \text{ for some } \theta \in \Theta \text{ with } \theta = \theta_{[b]}\}. \tag{2.4}$$

As defined, $\mathscr{Z}^0$ is the set of pairs of model and moment selection vectors (b, c) in $\mathscr{S}$ that select *only* moments that equal zero asymptotically for some $\theta \in \Theta$ with $\theta = \theta_{[b]}$. (The notation '$\mathscr{Z}^0$' is meant to remind one of 'zero under P^0'.) Define

$$\mathscr{M}\mathscr{Z}^0 = \{(b,\ c) \in \mathscr{Z}^0\colon\ |c| - |b| \geqslant |c^*| - |b^*|\ \forall (b^*, c^*) \in \mathscr{Z}^0\}. \tag{2.5}$$

As defined, $\mathscr{M}\mathscr{Z}^0$ is the set of selection vectors in $\mathscr{Z}^0$ that maximize the number of over-identifying restrictions out of the model and moment selection vectors in $\mathscr{Z}^0$. (The notation '$\mathscr{M}\mathscr{Z}^0$' denotes 'maximal over-identifying restrictions under P^0'.)

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

Assumption IDbc. $\mathscr{M}\mathscr{Z}^0$ contains a single element $(b^0,\ c^0)$.

When Assumption ID*bc* holds, we call b^0 the 'correct' model selection vector and c^0 the 'correct' moment selection vector. The correct selection vectors $(b^0,\ c^0)$ have the property that they uniquely select the maximal number of over-identifying restrictions out of all possible models and moment conditions. Depending upon P^0, Assumption ID*bc* may or may not hold. Below we analyze the properties of model and moment selection procedures both when this identification assumption holds and when it fails to hold.

When the maximum number of over-identifying restrictions is zero for any model and any set of moment conditions, i.e., $|c| - |b| \leqslant 0$ for $(b,\ c) \in \mathscr{M}\mathscr{Z}^0$, then Assumption ID*bc* typically does not hold. The reason is that whenever there are as many or more parameters $|b|$ as moment conditions $|c|$ there is usually some $|b|$-vector $\theta_{[b]} \in \Theta$ that solves the $|c|$ moment conditions $G_c(\theta_{[b]}) = \mathbf{0}$. Hence, $\mathscr{Z}^0$ typically contains multiple elements with $|c| = |b|$. In consequence, Assumption ID*bc* typically requires one or more *over-identifying* restrictions for it to hold. That is, it requires $|c| > |b|$ for $(b,\ c) \in \mathscr{M}\mathscr{Z}^0$.

For the model corresponding to the model selection vector b, let $\Theta_{[b]}(\subseteq \Theta)$ denote the parameter space. By definition, $\Theta_{[b]}$ is the subset of vectors in Θ that have zeros for elements that correspond to the zeros in b.

For distributions P^0 for which Assumption IDbc holds, we consider the following condition:

Assumption ID θ. There is a unique vector $\theta^0 \in \Theta_{[b^0]}$ such that $G^0_{c^0}(\theta^0) = \mathbf{0}$.

When Assumption IDθ 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 in $\Theta_{[b^0]}$ that does so.

Note that the standard GMM situation considered in the literature corresponds to the case where $\mathscr{MZ}^0 = \{(\mathbf{1}_p, \mathbf{1}_r)\}$ and Assumption IDθ is imposed, where $\mathbf{1}_p$ and $\mathbf{1}_r$ denote p- and r-vectors of ones. In this case, Assumption IDbc holds.

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

Next, we discuss Assumptions IDbc and IDθ in the context of linear IV estimation. Consider the iid linear regression model $Y_i = X_i'\theta^* + U_i$ for $i = 1, \ldots, n$ under P^0, where $E^0 U_i = 0$ and $E^0\|X_i\| < \infty$. We consider the IVs $\tilde{Z}_i \in R^r$, where $A^0 = E^0\tilde{Z}_i X_i' \in R^{r \times p}$ and $\rho^0 = E^0\tilde{Z}_i U_i \in R^r$. The moments in this case are $G_n(\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\theta)\tilde{Z}_i$ and the corresponding limit function is $G^0(\theta) = E^0(Y_i - X_i'\theta)\tilde{Z}_i = \rho^0 - A^0(\theta - \theta^*)$.

Let b^* $(\in R^p)$ denote the selection vector that selects all of the elements of θ^* that are not equal to zero. That is, the jth element of b^* is one if the corresponding element of θ^* is non-zero and is zero otherwise. Let c^* $(\in R^r)$ denote the selection vector that selects all of the IVs that are not correlated with the error U_i. Thus, the jth 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 in the correct model, i.e., $|c^*| > |b^*|$. In this context, the correct selection vector of regressors that enter the model is b^*, the selection vector of correct IVs is c^*, and the parameter of interest is $\theta^*_{[b^*]}$.

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

We now return to the general case. If Assumption IDbc fails to hold for some P^0, then it is still possible to define a 'correct' vector (b^0, c^0) in some cases. For given $P^0 \in \mathscr{P}$, we consider the following assumption:

Assumption IDbc2. $\mathscr{MZ}^0$ contains a single element (b^0, c^0) for which $|b^0| = \min\{|b|: (b, c) \in \mathscr{MZ}^0\}$.

That is, if it exists, we can define (b^0, c^0) to be the unique selection vector that provides the smallest parameterization of the model out of all selection vectors that maximize the number of over-identifying restrictions. Depending upon the circumstances, this may or may not be a suitable way of defining (b^0, c^0). Below, we focus on the definition of (b^0, c^0) given in Assumption ID*bc*, but we indicate results that apply when (b^0, c^0) is defined more generally by Assumption ID*bc*2.

2.3. The J-test statistic

All of the model and moment selection procedures considered below are based on the J-test statistic used for testing over-identifying restrictions, see Hansen (1982). We define this statistic here. The J-test statistic based on the model selected by b and the moments selected by c is defined to be

$$J_n(b, c) = n \inf_{\theta_{[b]} \in \Theta_{[b]}} G_{nc}(\theta_{[b]})' W_n(b, c) G_{nc}(\theta_{[b]}). \tag{2.6}$$

Here, $W_n(b, c)$ is the $|c| \times |c|$ weight matrix employed with the moments $G_{nc}(\theta_{[b]})$ and the model selected by b. For example, $W_n(b, c)$ might be defined such that it is an asymptotically optimal weight matrix when the moments selected by c are correct.[2] By definition, when $c = \mathbf{0}$, $W_n(b, c) = 0$ $(\in R)$.

The GMM estimator based on the model selected by b and the moments selected by c is defined to be any vector $\hat{\theta}_n(b, c) \in \Theta_{[b]}$ for which

$$G_{nc}(\hat{\theta}_n(b, c))' W_n(b, c) G_{nc}(\hat{\theta}_n(b, c)) = \inf_{\theta \in \Theta_{[b]}} G_{nc}(\theta)' W_n(b, c) G_{nc}(\theta). \tag{2.7}$$

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

$$V_n(b,c) = \frac{1}{n}\sum_{i=1}^{n} (m_c(Z_i, \tilde{\theta}_n(b, c)) - \bar{m}_{nc}(\tilde{\theta}_n(b, c)))(m_c(Z_i, \tilde{\theta}_n(b,c)) - \bar{m}_{nc}(\tilde{\theta}_n(b,c)))',$$

where $\bar{m}_{nc}(\theta) = (1/n)\sum_{i=1}^n m_c(Z_i, \theta)$ and $\tilde{\theta}_n(b,c)$ 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.

Thus, the $J_n(b, c)$ test statistic also can be written as

$$J_n(b, c) = nG_{nc}(\hat{\theta}_n(b, c))'W_n(b, c)G_{nc}(\hat{\theta}_n(b, c)). \tag{2.8}$$

2.4. The parameter space for the model and moment selection vectors

We consider estimation of (b^0, c^0) via an estimator that we denote generically by $(\hat{b}, \hat{c})$. The parameter space for $(\hat{b}, \hat{c})$ is denoted by $\mathscr{BC} \subset \mathscr{S}$. We always specify the parameter space $\mathscr{BC}$ such that it includes some $(b, c) \in \mathscr{S}$ with $c = \mathbf{0}$. This guarantees that the parameter space always includes at least one pair (b, c) of selection vectors that does not select any incorrect moments (since it does not select any moments at all). Note that the lack of any correct moments indicates model misspecification.

The parameter space $\mathscr{BC}$ should be a very much smaller set than $\mathscr{S}$. Otherwise, the finite sample behavior of $(\hat{b}, \hat{c})$ will be poor and computation will be difficult. The parameter space $\mathscr{BC}$ should exploit the information that many parameters are known not to be zero and that many moment conditions are known to be correct. It should also exploit the nested or hierarchical structure that often exists with parameters (e.g., with lagged variables) and with moment conditions (e.g., when blocks of moment conditions are either correct or incorrect block by block rather than moment condition by moment condition, see Andrews (1999)).

2.5. Definition of consistency

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 $(\hat{b}, \hat{c}) \in \mathscr{BC}$ is *consistent* if

$$(\hat{b}, \hat{c}) = (b^0, c^0) \text{ wp} \to 1 \text{ under } P^0, \ \forall P^0 \in \mathscr{P} \text{ that satisfy Assumption ID}bc. \tag{2.9}$$

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

We note that the above definition of consistency is stronger if Assumption ID*bc* is replaced by Assumption ID*bc*2 in (2.9).

2.6. Performance when assumption IDbc fails

Below we analyze the behavior of the model and moment selection procedures introduced below in the case where Assumption ID*bc* does not hold. For this purpose, we make the following definitions. Define

$$\mathscr{BCZ}^0 = \mathscr{BC} \cap \mathscr{Z}^0. \tag{2.10}$$

As defined, $\mathscr{BCZ}^0$ is the set of selection vectors in the parameter space $\mathscr{BC}$ that select only models and moments that equal zero asymptotically for some parameter vector. Define

$$\mathscr{MBCZ}^0 = \{(b, c) \in \mathscr{BCZ}^0\colon |c| - |b| \geqslant |c^*| - |b^*| \ \forall (b^*, c^*) \in \mathscr{BCZ}^0\}. \tag{2.11}$$

As defined, $\mathscr{MBCZ}^0$ is the set of selection vectors in $\mathscr{BCZ}^0$ that maximize the number of over-identifying restrictions out of selection vectors in $\mathscr{BCZ}^0$. We show below that for many moment selection procedures discussed below $(\hat{b}, \hat{c}) \in \mathscr{MBCZ}^0$ wp $\to 1$ whether or not Assumption ID*bc* holds. That is, for these procedures, with probability that goes to one as $n \to \infty$, $(\hat{b}, \hat{c})$ lies in the set of selection vectors that *maximize* the number of over-identifying restrictions out of all selection vectors in the parameter space $\mathscr{BC}$ that select only moments that equal *zero* asymptotically for some parameter vector.

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, $\forall P^0 \in \mathscr{P}$.

(b) $W_n(b, c) \to_p W^0(b, c)$ under P_0 for some positive definite matrix $W^0(b, c)$ $\forall (b, c) \in \mathscr{BC}$, $\forall P_0 \in \mathscr{P}$.

(c) $\inf_{\theta \in \Theta_{[b]}} G_{nc}(\theta)' W_n(b, c) G_{nc}(\theta) \to_p \inf_{\theta \in \Theta_{[b]}} G_c^0(\theta)' W^0(b, c) G_c^0(\theta) = G_c^0(\theta^*)' W^0(b, c) \times G_c^0(\theta^*)$ under P^0 for some $\theta^* \in \Theta_b$ that may depend on c and P^0, $\forall (b, c) \in \mathscr{BC}$, $\forall P^0 \in \mathscr{P}$.

Assumption 1(a) typically holds by a central limit theorem (CLT) with $G^0(\theta)$ equal to the expectation of $G_n(\theta)$ or its limit as $n \to \infty$, 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 $W_n(b, c)$.

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 $G^0(\cdot)$ as in Assumption 1(a), $G^0(\theta)$ is continuous on Θ, and $\Theta_{[b]} \subset R^p$ is compact for all b such that $(b, c) \in \mathscr{BC}$ for some c, $\forall P^0 \in \mathscr{P}$. The first two of these three conditions can be verified using a generic uniform convergence result, such as a uniform weak LLN, e.g., see Andrews (1992). Alternatively, when the moments are linear in θ, Assumption 1(c) typically holds under almost the same conditions as Assumption 1(a), because the ‘infima over $\theta \in \Theta_{[b]}$’ can be calculated explicitly. In the linear case, the parameter spaces $\Theta_{[b]}$ can be unbounded.

For illustrative purposes, we provide a sufficient condition for Assumption 1 for the case of stationary data. This condition is not very restrictive. (The proof of sufficiency is given in Andrews (1999).) Let $\|B\|$ denote the Euclidean norm of a vector or matrix, i.e., $\|B\| = (\operatorname{tr} B'B)^{1/2}$.

Assumption STAT. (a) $\{Z_i: i = \ldots, 0, 1, \ldots\}$ is a doubly infinite stationary and ergodic sequence under P^0, $\forall P^0 \in \mathscr{P}$.

(b) $G_n(\theta) = (1/n)\sum_{i=1}^{n} m(Z_i, \theta)$ 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 $\sum_{j=1}^{\infty} (E^0\|E^0(m(Z_i, \theta)|\mathscr{F}_{i-j})\|^2)^{1/2} < \infty$ $\forall \theta \in \Theta$, $\forall P^0 \in \mathscr{P}$, where $\mathscr{F}_i$ denotes the σ-field generated by $(\ldots, Z_{i-1}, Z_i)$.

(d) Either (i) $\Theta_{[b]} \subset R^p$ is compact for all b such that $(b, c) \in \mathscr{BC}$ for some c and $E^0 \sup_{\theta\in\Theta} \|m(Z_i, \theta)\| < \infty$ $\forall P^0 \in \mathscr{P}$ 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 $\Theta_{[b]} = \{\theta \odot b: \theta \in R^p\}$ for all b such that $(b, c) \in \mathscr{BC}$ for some c, *where* '$\odot$' *denotes element by element product.*

(e) *Assumption 1* (b) *holds.*

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

3. Model and moment selection criteria

Here we introduce a class of model and moment selection criteria (MMSC) that are analogous to the well-known model selection criteria used for choosing between competing models. They extend the moment selection criteria of Andrews (1999) to allow for simultaneous model and moment selection.

The MMSC estimator, $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}})$, is the value that minimizes

$$\mathrm{MMSC}_n(b, c) = J_n(b, c) - h(|c| - |b|)\kappa_n \tag{3.1}$$

over $\mathscr{BC}$. The function $h(\cdot)$ and the constants $\{\kappa_n: n \geqslant 1\}$ in the definition of $\mathrm{MMSC}_n(b, c)$ are specified by the researcher. They are assumed to satisfy:

Assumption MMSC. (a) $h(\cdot)$ is strictly increasing.

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

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

It is always possible to specify MMSC for which Assumption MMSC holds, because the researcher chooses $h(\cdot)$ and $\{\kappa_n\colon n \geqslant 1\}$.

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

$$\begin{aligned}
&\text{MMSC–BIC: } \kappa_n = \ln n \text{ and } \mathrm{MMSC}_{\mathrm{BIC},n}(b, c) = J_n(b, c) - (|c| - |b|)\ln n,\\
&\text{MMSC–AIC: } \kappa_n = 2 \text{ and } \mathrm{MMSC}_{\mathrm{AIC},n}(b, c) = J_n(b, c) - 2(|c| - |b|),\\
&\text{MMSC–HQIC: } \kappa_n = Q\ln\ln n \text{ for some } Q > 2 \text{ and}\\
&\qquad\qquad \mathrm{MMSC}_{\mathrm{HQIC},n}(b, c) = J_n(b, c) - Q(|c| - |b|)\ln\ln n. \qquad (3.2)
\end{aligned}$$

The MMSC–BIC and MMSC–HQIC procedures satisfy Assumption MMSC. The MMSC–AIC procedure does not satisfy Assumption MMSC(b) because $\kappa_n = 2 \nrightarrow \infty$. Thus, the MMSC–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 MMSC–AIC procedure has positive probability even asymptotically of selecting too few over-identifying restrictions.

Consistency of $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}})$ is established in the following theorem.

Theorem 1. Suppose Assumptions 1 and MMSC *hold. Then,*

(a) $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}}) \in \mathscr{MBCZ}^0$ wp $\to 1$, $\forall P^0 \in \mathscr{P}$,

(b) *for all* $P^0 \in \mathscr{P}$ *for which Assumption* IDbc *holds*, $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}}) = (b^0, c^0)$ wp $\to 1$ *iff* $(b^0, c^0) \in \mathscr{BC}$, *and*

(c) $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}})$ *is consistent iff for all* $P^0 \in \mathscr{P}$ *for which Assumption* IDbc *holds, we have* $(b^0, c^0) \in \mathscr{BC}$.

Comment. 1. Part (a) is a robust result that specifies the asymptotic behavior of $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}})$ for all $P^0 \in \mathscr{P}$, not just for P^0 for which Assumption IDbc holds. Note that if $\mathscr{MBCZ}^0 \cap \mathscr{MZ}^0 \neq \emptyset$, then $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}}) \in \mathscr{MZ}^0$ wp $\to 1$, $\forall P^0 \in \mathscr{P}$. The result of part (a) is analogous to results concerning the behavior of extremum estimators when the standard identification condition fails.

2. Theorem 1 is analogous to Theorem 1 of Andrews (1999). Theorem 1(b) is similar to Theorem 3 of Hannan (1980) for (weak) consistency of model selection criteria for lag selection in ARMA models.

3. Over-rejection of the J test in finite samples (see the July 1996 issue of the *Journal of Business and Economic Statistics*) affects the MMSC only if the amount of over-rejection differs for different selection vectors (b, c).

4. The proof of Theorem 1 is given in the Appendix of Proofs.

5. Suppose that consistency is defined with Assumption IDbc replaced by Assumption ID$bc$2. Then, a consistent MMSC can be obtained by adding a penalty term $h_2(|b|)\kappa_{2n}$ to the definition of $\mathrm{MMSC}_n(b, c)$ in (3.1), where $h_2(\cdot)$ is a strictly increasing function, $\kappa_{2n} \to \infty$, and $\kappa_{2n} = o(\kappa_n)$.

4. Downward testing procedure

The downward testing (DT) procedure considered in this section is a model and moment selection procedure that formalizes the procedure that empirical researchers often use in a less formal fashion. Two advantages of considering a precisely specified model and moment selection procedure are that (i) sufficient conditions for consistency can be established and (ii) the effect of selection on post-selection statistical inference can be assessed, e.g., via simulations.

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

Note that, for a given number of moments, the downward testing model and moment selection procedure progresses from the most restrictive model to the least restrictive. 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 model and moment selection procedures, are referenced in Amemiya (1980) and Pötscher (1989).

We now define $\hat{k}_{\mathrm{DT}}$ and $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}})$ more precisely. Let $\gamma_{n,k} > 0$ denote the critical value employed with the test statistic $J_n(b, c)$ when $|c| - |b| = k$ and the sample size is n. In the recommended case where $J_n(b, c)$ is constructed using an asymptotically optimal weight matrix, $J_n(b, c)$ has an asymptotic chi-square distribution with $|c| - \min(|b|, |c|)$ degrees of freedom when all moment conditions in c are correct given the model selected by b.[3] In this case,

[3] 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.

one takes

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

for values of $k > 0$, where $\chi^2_k(\alpha_n)$ denotes $1 - \alpha_n$ quantile of a chi-squared distribution with k degrees of freedom.

Let $\hat{k}_{\mathrm{DT}} \in \{-p, -p+1, \dots, r\}$ be such that $\min_{(b,\, c) \in \mathscr{BC}:\ |c|-|b|=k} J_n(b, c) > \gamma_{n,k}\ \forall k > \hat{k}_{\mathrm{DT}}$ with $k \in \mathscr{K} = \{|c| - |b|\colon (b, c) \in \mathscr{BC}\}$, $\min_{(b,c)\in\mathscr{BC}:|c|-|b|=\hat{k}_{\mathrm{DT}}} J_n(b,c) \leqslant \gamma_{n,\hat{k}_{\mathrm{DT}}}$, and $\hat{k}_{\mathrm{DT}} \in \mathscr{K}$. Define $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}})$ to be any vector in $\mathscr{BC}$ for which $|\hat{c}_{\mathrm{DT}}| - |\hat{b}_{\mathrm{DT}}| = \hat{k}_{\mathrm{DT}}$ and $J_n(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}}) = \min_{(b,c)\in\mathscr{BC}:\ |c|-|b|=\hat{k}_{\mathrm{DT}}} J_n(b, c)$. In words, $\hat{k}_{\mathrm{DT}}$ is the greatest number of over-identifying restrictions for which some $J_n(b, c)$ test does not reject for some $(b, c) \in \mathscr{BC}$. Given $\hat{k}_{\mathrm{DT}}$, $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}})$ is the vector that minimizes $J_n(b, c)$ over vectors $(b, c) \in \mathscr{BC}$ with $|c| - |b| = \hat{k}_{\mathrm{DT}}$.

For consistency of $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}})$, we assume the critical values $\gamma_{n,k}$ satisfy:

Assumption T. $\gamma_{n,k} \to \infty$ and $\gamma_{n,k} = \mathrm{o}(n)\ \forall k \in \mathscr{K}$.

Assumption T holds if $\{\gamma_{n,k}\colon k \in \mathscr{K}\}$ are defined as in (4.1) with the significance level α_n satisfying $\alpha_n \to 0$ and $\ln \alpha_n = \mathrm{o}(n)$ (see Theorem 5.8 of Pötscher (1983)). For example, the latter condition holds if $\alpha_n \geqslant \lambda_0 \exp(-\lambda_n n)$, for some $0 < \lambda_n \to 0$ and $\lambda_0 > 0$.

Consistency of $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}})$ is established in the following theorem.

Theorem 2. Suppose Assumptions 1 and T *hold. Then,*

(a) $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}}) \in \mathscr{MBCZ}^0$ wp → 1, $\forall P^0 \in \mathscr{P}$,

(b) *for all* $P^0 \in \mathscr{P}$ *for which Assumption IDbc holds,* $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}}) = (b^0, c^0)$ wp → 1 *iff* $(b^0, c^0) \in \mathscr{BC}$, *and*

(c) $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}})$ *is consistent iff for all* $P^0 \in \mathscr{P}$ *for which Assumption IDbc holds, we have* $(b^0, c^0) \in \mathscr{BC}$.

Comment. 1. Theorem 2 is similar to Theorem 2 of Andrews (1999) for consistency of downward testing moment selection procedures and Theorem 5.7 of Pötscher for consistency of upward LM tests for lag selection in ARMA models.

2. The testing procedure $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}})$ determines when there are no over-identifying restrictions, just as $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}})$ does.

3. 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 over-identifying restrictions, but not necessarily all of them.

4. One can also consider upward testing procedures, as in Andrews (1999). These procedures have the drawback that they are consistent only under an additional restriction, see Andrews (1999). For this reason and for brevity, we do not consider upward testing procedures explicitly here.

5. An application to dynamic panel data models

5.1. A general dynamic model for panel data

Consider a dynamic panel data model

$$\begin{aligned} y_{it} &= w_{it}'\delta_t + u_{it}, \\ u_{it} &= \eta_i + v_{it}, \quad \forall t = 1,\ldots,T \text{ and } i = 1,\ldots,N. \end{aligned} \tag{5.1}$$

Here y_{it} and w_{it} are observed variables, v_{it} is an unobserved idiosyncratic error, η_i is an unobserved individual effect, and δ_t are unknown parameters to be estimated. The distributions of η_i and v_{it} are not specified, but assumptions on their means and correlations with other variables are given below. All of the random variables in the model are assumed to be independent across individuals i.

The regressor vector w_{it} includes L lags of the dependent random variable, i.e., $y_{it-1},\ldots,y_{it-L}$, where $L \geqslant 0$. The true lag length L_0 ($\leqslant L$) may be unknown. The initial observations $\{y_{i0}, y_{i,-1},\ldots,y_{i,1-L}$: $i = 1, 2,\ldots,N\}$ are assumed to be observed.

The regressor vector w_{it} also includes other variables that may be strictly exogenous, predetermined, or endogenous. These other variables are contained in two observed vectors z_{it} and f_i of time varying and time invariant variables respectively. The vectors z_{it} and f_i may also contain variables that do not enter the regression function. Such variables can be employed as instrumental variables.

The time varying variables z_{it} (and, hence, the time varying regressors in w_{it}) may consist of five types of variables. The type of a variable depends on whether the variable is strictly exogenous, predetermined, or endogenous with respect to v_{it} and whether the variable is uncorrelated or correlated with the individual effect η_i. We partition z_{it} as

$$z_{it} = (x_{1it}', x_{2it}', p_{1it}', p_{2it}', y_{2it}')'. \tag{5.2}$$

Here, the variables $(x_{1it}', x_{2it}')'$ are strictly exogenous with respect to v_{it}. The variables $(p_{1it}', p_{2it}')'$ are predetermined with respect to v_{it}. The variables y_{2it} are endogenous with respect to v_{it}. The variables $(x_{1it}', p_{1it}')'$ are uncorrelated with the individual effect η_i. The variables $(x_{2it}', p_{2it}', y_{2it}')'$ are correlated with the individual effect η_i. The econometrician may not know the type of some variables in z_{it} (and, hence, of some regressors in w_{it}).

The time invariant variables f_i (and, hence, the time invariant regressors in w_{it}) are strictly exogenous with respect to v_{it} and of two types. The type of a variable depends on whether the variable is uncorrelated or correlated with the individual effect η_i. We partition f_i as

$$f_i = (f_{1i}', f_{2i}')'. \tag{5.3}$$

Here, the variables f_{1i} are uncorrelated with the individual effect η_i and the variables f_{2i} are correlated with η_i. The econometrician may not know whether certain variables in f_i are uncorrelated or correlated with η_i.

If z_{it} and f_i do not contain any variables other than those that enter the vector of regressors w_{it}, then general model (5.1) can be written as

$$y_{it} = \sum_{m=1}^{L} \alpha_{mt} y_{i,t-m} + z_{it}'\beta_t + f_i'\gamma_t + u_{it},$$

$$u_{it} = \eta_i + v_{it} \quad \forall t = 1, \ldots, T \text{ and } i = 1, \ldots, N, \tag{5.4}$$

where $w_{it} = (y_{i,t-1}, \ldots, y_{i,t-L}, z_{it}', f_i')'$ and $\delta_t = (\alpha_{1t}, \ldots, \alpha_{tL}, \beta_t', \gamma_t')'$. (Note that this model includes intercept parameters provided f_i contains a constant.)

In the general model (5.1), the parameter δ_t can vary with t. For example, this allows one to consider a model with structural breaks at known or unknown times. If a structural break occurs, the parameter takes different values before and after the break point. To conform with the set-up of Section 2, we parameterize the model in terms of the parameter values for the first period, denoted δ, and the corresponding deviations from these values, denoted $\delta_{(t)}$ for $t = 2, \ldots, T$:

$$\delta_t = \delta + \delta_{(t)}, \tag{5.5}$$

In the most general case, (5.5) allows $\delta_{(t)}$ (and δ_t) to take different values for each t and the parameter vector θ is defined to be

$$\theta = (\delta', \delta_{(2)}', \ldots, \delta_{(T)})'. \tag{5.6}$$

Usually in practice, however, one will use a restricted version of (5.5), which leads to a ‘restricted model’ rather the fully general model (5.1).

Examples of restricted models are: (i) No structural breaks occur over the sample period, i.e., $\delta_{(t)} = \mathbf{0}$ $\forall t = 2, \ldots, T$. In this case, the parameter vector θ simplifies to

$$\theta = \delta. \tag{5.7}$$

(ii) H structural breaks occur at times $1 < \tau_1 < \tau_2 < \cdots < \tau_H \leqslant T$. Then,

$$\delta_{(t)} = \delta^{(k)} \quad \forall t \text{ with } \tau_k \leqslant t < \tau_{k+1}, \quad k = 1, \ldots, H, \tag{5.8}$$

where $\tau_{H+1} = T + 1$. In this case, the parameter vector θ simplifies to

$$\theta = (\delta', \delta^{(1)'}, \ldots, \delta^{(H)'})'. \tag{5.9}$$

(iii) H or fewer structural breaks occur at unknown times. For each combination of a number of structural breaks and times of the breaks that is to be considered, one specifies vectors of ‘deviation’ parameters as in (5.8). Then, the first period parameter δ and all of the deviation parameters are stacked into a single vector θ. By appropriately selecting different subsets of the deviation parameters, one obtains models with different numbers and times of structural change.

(iv) Partial structural breaks occur. In this case, structural breaks occur at known or unknown times, but only a subset of the parameters δ_t change. For brevity, we do not provide the details.

It is worth mentioning that structural breaks in the individual effect also could be introduced in the general model by allowing η_i to have different coefficients for different time periods, as in Chamberlain (1984) and Holtz-Eakin et al. (1988). We do not do so here, because this would lead to different moment conditions than those considered in most dynamic panel data models considered in the literature.

To this point, we have kept a high level of generality in model (5.1) by incorporating many features that arise in different empirical studies. For example, allowing for an unknown lag length is especially important for purely dynamic panel data models that do not have any other regressors. Whether elements of w_{it} are predetermined or strictly exogenous with respect to v_{it} is especially important in models with rational expectations. Whether variables in $(z'_{it}, f'_i)'$ are correlated with the individual effect or not separates the 'correlated random effects' model from the standard 'random effects' model and is important for many applications. Allowing for structural breaks in the parameters provides a way to model nonstationarity in dynamic panel data models that is an alternative to panel data unit root models. It is useful in many applications.

On the other hand, we do not expect that in any particular empirical study, all of these features will be present or important simultaneously. The purpose of the generality of model (5.1) is to have a single theoretical framework that covers a wide variety of more restrictive sub-models that are of interest in different applications.

5.2. Comparison with panel data models in the literature

Here we show that the general model (5.1) nests many models in the literature and shares common features with some others.

Model (5.1) becomes the standard static 'random effects' model, if there are no lagged dependent variables in the model, i.e., $L = 0$, all of the regressors w_{it} are strictly exogenous with respect to v_{it}, none of the regressors w_{it} are correlated with η_i, and the parameters are constant over time.

The following static *correlated* random effects model is considered by Hausman and Taylor (1981) and Breusch et al. (1989):

$$\begin{aligned} y_{it} &= z'_{it}\beta + f'_i\gamma + u_{it}, \\ u_{it} &= \eta_i + v_{it} \quad \forall t = 1, \ldots, T \text{ and } i = 1, \ldots, N. \end{aligned} \tag{5.10}$$

This model does not contain any lagged values of y_{it}. In our notation, the regressor vector w_{it} equals $(z'_{it}, f'_i)'$. The regressors z_{it} and f_i are assumed to be strictly exogenous with respect to v_{it} and a subset of z_{it} and f_i are correlated with the individual effect η_i. That is, in our notation, $z_{it} = x_{it} = (x'_{1it}, x'_{2it})'$, and $f_i = (f'_{1i}, f'_{2i})'$. This model also is one of four models considered in Amemiya and MaCurdy (1986). All of the authors above consider estimation of this model by instrumental variables.

Anderson and Hsiao (1982) and Bhargava and Sargan (1983) consider maximum likelihood estimation of a dynamic panel data model

$$\begin{aligned} y_{it} &= \alpha_1 y_{i,t-1} + z'_{it}\beta + f'_i\gamma + u_{it}, \\ u_{it} &= \eta_i + v_{it} \quad \forall t = 1, \ldots, T \text{ and } i = 1, \ldots, N. \end{aligned} \tag{5.11}$$

They also consider simpler versions of this model. Here, both z_{it} and f_i are assumed to be strictly exogenous with respect to v_{it} and uncorrelated with η_i. In our notation, $w_{it} = (y_{i,t-1}, z'_{it}, f'_i)'$, $z_{it} = x_{1it}$, and $f_i = f_{1i}$. The lag length of the lagged dependent variables is known to be one. These authors assume normal distributions for η_i and v_{it}. Because the number of time series observations T is small for typical panels, the assumption used by these authors concerning the initial observation plays a crucial role in interpreting the model and obtaining a consistent estimator. These authors also discuss the case where v_{it} is serially correlated. We do not consider this case in the present paper.

Ahn and Schmidt (1995) consider GMM estimation of several dynamic and static panel data models. The most general model they consider is

$$\begin{aligned} y_{it} &= \alpha_1 y_{i,t-1} + z'_{it}\beta + f'_i\gamma + u_{it}, \\ u_{it} &= \eta_i + v_{it} \quad \forall t = 1, \ldots, T \text{ and } i = 1, \ldots, N. \end{aligned} \tag{5.12}$$

This model contains only one lagged value of y_{it}. The regressors z_{it} and f_i are assumed to be strictly exogenous with respect to v_{it} and a subset of z_{it} and f_i may be correlated with the individual effect η_i. That is, in our notation, $w_{it} = (y_{i,t-1}, z'_{it}, f'_i)'$, $z_{it} = x_{it} = (x'_{1it}, x'_{2it})'$ and $f_i = (f'_{1i}, f'_{2i})'$.

Arellano and Bover (1995) consider a model that nests models (5.10)–(5.12). It allows the regressor vector w_{it} and z_{it} to contain strictly exogenous, predetermined, and endogenous variables with respect to v_{it}. That is, as in our model (5.1), $z_{it} = (x'_{1it}, x'_{2it}, p'_{1it}, p'_{2it}, y'_{2it})'$. Our model (5.1) nests that of Arellano and Bover (1995) and models (5.10)–(5.12) in that it allows for time-varying parameters.

Holtz-Eakin et al. (1988) consider a bivariate vector autoregression (VAR) of $(y_{it}, \tilde{y}_{it})$ with panel data. In our notation, each equation of their VAR takes the form

$$y_{it} = \sum_{m=1}^{L_0} \alpha_{mt} y_{i,t-m} + z'_{it}\beta_t + \gamma_t + u_{it},$$

$$u_{it} = \lambda_t \eta_i + v_{it} \quad \forall t = 1,\ldots,T \text{ and } i = 1,\ldots,N. \tag{5.13}$$

In this model, the true lag length L_0 of the lagged y_{it} variables is assumed to be known. The time varying regressors z_{it} contain only lagged values of the second endogenous variable $\tilde{y}_{it}$ and, thus, contain only variables that are predetermined with respect to v_{it} and are correlated with η_i. That is, in our notation, $z_{it} = p_{2it}$. Also, in their model, the only time-invariant strictly exogenous variable is a constant. Thus, in our notation, $f_i = 1$. These aspects of (5.13) are less general than corresponding parts of our model (5.1).

On the other hand, (5.13) allows for a time-varying coefficient λ_t on the individual effect. Model (5.13) is more general than (5.1) in this respect. Such generality comes at the expense of identification, however, because at best only the ratios λ_t/λ_{t-1} may be identified. Holtz-Eakin et al. (1988) do not provide identification results for their most general model (5.13), nor do they consider estimation of it, but they do provide tests for whether a more restrictive model with constant coefficients is sufficiently general.

5.3. Moment conditions in dynamic panel data models

It is well known that the simple OLS estimator of (5.1) is inconsistent because the lagged dependent variables $y_{i,t-1}, y_{i,t-2}, \ldots, y_{i,t-L}$ and (possibly unknown) subsets of other regressors are correlated with the unobserved individual effect η_i. In consequence, we consider GMM estimation of model (5.1).

The moment conditions that are employed by a GMM estimator are implied by assumptions that are imposed on the dynamic panel data model. Below, we state various assumptions and corresponding moment conditions that can be used with model (5.1). We state the assumptions sequentially such that they impose increasingly restrictive assumptions on the model. The use of different combinations of the assumptions yields different models. We do not require that all of the assumptions are imposed.

We note that the use of additional correct moment conditions can substantially improve the efficiency of an estimator in some cases; e.g., see Blundell and Bond (1995). Furthermore, the identification of some parameters and the consistency of an estimator may rely on some moment conditions being correct and being employed by the estimator. On the other hand, the use of incorrect moment conditions typically leads to inconsistency of an estimator.

In what follows, we use the notation $z_{it} = (x'_{1it}, x'_{2it}, p'_{1it}, p'_{2it}, y_{2it})'$, $f_i = (f'_{1i}, f'_{2i})'$, $x_{it} = (x'_{1it}, x'_{2it})'$, and $p_{it} = (p'_{1it}, p'_{2it})'$. Each assumption applies for all $i = 1,\ldots,N$.

Assumption P1. (a) $\mathrm{E}\eta_i = 0$, $\mathrm{E}v_{it} = 0$, $\mathrm{E}v_{it}\eta_i = 0\ \forall t = 1, 2, \ldots, T$.
(b) $\mathrm{E}v_{is}v_{it} = 0\ \forall s, t = 1, 2, \ldots, T$ with $s \neq t$.
(c) $\mathrm{E}v_{it}y_{i0} = \cdots = \mathrm{E}v_{it}y_{i,1-L} = 0\ \forall t = 1, 2, \ldots, T$.
(d) $\mathrm{E}v_{it}(z'_{i1}, \ldots, z'_{it-1}, x'_{it}, p'_{it}, f'_i)' = 0\ \forall t = 1, 2, \ldots, T$.

Assumption P2. $\mathrm{E}v_{it}(x'_{i,t+1}, \ldots, x'_{iT})' = 0\ \forall t = 1, 2, \ldots, T-1$.

Assumption P3. $\mathrm{E}\eta_i(x'_{1it}, p'_{1it}, f'_{1i})' = 0\ \forall t = 1, 2, \ldots, T$.

Assumption P4. $\mathrm{E}\eta_i(x'_{2it}, p'_{2it}, y'_{2it})' = \mathrm{E}\eta_i(x'_{2i,t-1}, p'_{2i,t-1}, y'_{2i,t-1})'\ \forall t = 2, \ldots, T$.

Assumption P5. $\mathrm{Var}(v_{it}) = \sigma_i^2$ for some $\sigma_i^2 > 0\ \forall t = 1, 2, \ldots, T$.

Assumption P6. $\mathrm{E}\eta_i y_{i1} = \mathrm{E}\eta_i y_{it}\ \forall t = 1 - L, \ldots, 0$.

Assumptions P1(a)–(c) impose the familiar error-components structure and are referred to as 'standard assumptions' by Ahn and Schmidt (1995) for dynamic panel data models with only lagged dependent variables as regressors. Assumption P1(a) requires that the error u_{it} ($= v_{it} + \eta_i$) has mean zero and v_{it} is uncorrelated with the individual effect η_i. Assumption P1(b) requires that v_{it} is serially uncorrelated. Assumption P1(c) requires that v_{it} is uncorrelated with the initial observations. Assumption P1(d) requires that all lags of z_{it} are uncorrelated with v_{it} and that all variables in z_{it} except the endogenous variables y_{2it} are at least predetermined with respect to v_{it} (i.e., their current period correlation with v_{it} is also zero). Assumptions P1(a)–(d) are the minimum restrictions imposed on model (5.1). They may not identify γ_t, the coefficients on the time invariant regressors f_i.

Assumption P2 specifies that some of the variables in z_{it} are strictly exogenous with respect to v_{it}, rather than just predetermined.

Assumptions P3 and P4 concern the correlation between the regressors in $(z'_{it}, f'_i)'$ and the individual effect η_i. Assumption P3 specifies that some variables in z_{it} and f_i are uncorrelated with the individual effect η_i. This assumption can be used to identify γ_t. Assumption P4 specifies that the variables in z_{it} that are correlated with η_i have constant correlation across time with η_i. This type of restriction is considered by Bhargava and Sargan (1983) and Breusch et al. (1989).

Assumption P5 concerns the second moments of the error terms. It assumes that the variance of v_{it} is constant over time for each individual. (The variance may vary across individuals.) In the literature, Assumption P5 (plus the assumption that the variance of v_{it} is constant across individuals) is used to obtain a feasible GLS estimator for the random effects model and a 3SLS estimator for the correlated random effects model, because it implies a known structure for the variance–covariance matrix of the errors, which is needed for the GLS transformation. For a GMM estimator, the role of Assumption P5 is to provide additional moment conditions.

Assumption P6 is a 'stationarity' assumption on the initial conditions $y_{i,1-L}, \ldots, y_{i0}$. It requires that the initial conditions have same correlation with the individual effect as the dependent variable at time $t = 1$ has. The failure of this assumption indicates that $y_{i1-L}, \ldots, y_{i0}$ are not drawn from the same process that generates y_{i1}. Assumption P6 also is used by Arellano and Bover (1995), Blundell and Bond (1995), and Ahn and Schmidt (1995). Blundell and Bond (1995) study the usefulness of Assumption P6 via a Monte Carlo study of a simple dynamic panel data model with no regressors except a single lagged dependent variable. They show this assumption, if correct, can substantially improve the asymptotic efficiency of a GMM estimator when α_1, the coefficient on the lagged dependent variable, is close to unity.

We now specify the moment conditions that are implied by Assumptions P1–P6. Let Δ denote the first difference operator applied to the variable immediately following Δ. Thus, $\Delta u_{it} z_{it} = (u_{it} - u_{i,t-1}) z_{it}$.

Assumption P1 implies the following moment conditions:

$$\mathrm{E} u_{it} = 0 \quad \forall t = 1, \ldots, T, \tag{5.14}$$

$$\mathrm{E}\Delta u_{it}(y_{i,1-L}, \ldots, y_{i,t-2})' = 0 \quad \forall t = 2, \ldots, T, \tag{5.15}$$

$$\mathrm{E}\Delta u_{it}(z'_{i1}, \ldots, z'_{i,t-2}, x'_{i,t-1}, p'_{i,t-1}, f'_i)' = 0 \quad \forall t = 2, \ldots, T, \tag{5.16}$$

$$\mathrm{E} u_{it} \Delta u_{it-1} = 0 \quad \forall t = 3, \ldots, T. \tag{5.17}$$

Let d_z, d_f, d_x, and d_p denote the dimensions z_{it}, f_i, x_{it}, and p_{it}, respectively. The numbers of moment conditions in (5.14)–(5.17) are T, $L(T-1) + (T-2)(T-1)/2$, $d_z(T-1)(T-2)/2 + (d_x + d_p)(T-1) + d_f(T-1)$, and $T-2$ respectively.[4]

To construct a GMM estimator based on the moment conditions above (and those below), one replaces u_{it} by the difference between y_{it} and the regression function evaluated at the parameter vector θ (or $\theta_{[b]}$). Doing so, one can see that the moment conditions in (5.14)–(5.16) yield estimating equations that are linear in the parameters, whereas those generated by (5.17) are nonlinear in the parameters.

Assumption P2 implies the following moment conditions:

$$\mathrm{E}\Delta u_{it}(x'_{it}, \ldots, x'_{iT})' = 0 \quad \forall t = 2, \ldots, T. \tag{5.18}$$

The number of moment conditions in (5.18) is $d_x T(T-1)/2$.

[4] We note that an equivalent set of moment conditions to (5.14)–(5.17) are (5.14)–(5.16) plus $\mathrm{E} u_{iT} \Delta u_{it-1} = 0 \ \forall t = 3, \ldots, T$.

Assumption P3, combined with Assumption P1(d), implies the following moment conditions:

$$\begin{aligned}
&\mathrm{E}\mathbf{1}'_T u_i x_{1it} = 0 \quad \forall t = 1,\dots,T,\\
&\mathrm{E}(\mathbf{1}^T_{t+1})' u_i p_{1it} = 0 \quad \forall t = 1,\dots,T-1,\\
&E\mathbf{1}'_T u_i f_{1i} = 0,
\end{aligned} \tag{5.19}$$

where $u_i = (u_{i1},\dots,u_{iT})'$, $\mathbf{1}_T$ denotes a T-vector of ones, and $\mathbf{1}^T_{t+1}$ denotes a T-vector whose first t elements are zeros and whose elements indexed from $t+1$ to T equal ones. These moment conditions, if correct, can be used to identify γ_t. Let d_{x_1}, d_{p_1}, and d_{f_1} denote the dimensions of x_{1it}, p_{1it}, and f_{1i}, respectively, that are uncorrelated with the individual effect. The number of moment conditions in (5.19) is $d_{x_1}T + d_{p_1}(T-1) + d_{f_1}$.

Assumption P4, combined with Assumption P1(d), implies the following moment conditions:

$$\mathrm{E}u_{it}(\Delta x'_{2it}, \Delta p'_{2it}, \Delta y'_{2it})' = 0 \quad \forall t = 2,\dots,T. \tag{5.20}$$

Let d_{x_2}, d_{p_2}, and d_{y_2} denote the dimensions of x_{2it}, p_{2it}, and y_{2it}. The number of moment conditions in (5.20) is $(d_{x_2} + d_{p_2} + d_{y_2})(T-1)$.

Assumption P5 leads to the following $T-1$ moment conditions:

$$\mathrm{E}\mathbf{1}'_T u_i \Delta u_{it} = 0 \quad \forall t = 2,\dots,T. \tag{5.21}$$

Suppose one wishes to maximize the number of moment conditions that generate estimating equations that are linear in the parameters. Then, Ahn and Schmidt (1995) show that, when the homoskedasticity Assumption P5 holds, the moment conditions in (5.14)–(5.17) can be expressed equivalently as those in (5.14)–(5.16) plus

$$\mathrm{E}(y_{i,t-2}\Delta u_{i,t-1} - y_{i,t-1}\Delta u_{it}) = 0 \quad \forall t = 3,\dots,T. \tag{5.22}$$

Assumption P6 implies that

$$\mathrm{E}(\mathbf{1}^T_2)' u_i \Delta y_{i,t} = 0 \quad \forall t = 2-L,\dots,1. \tag{5.23}$$

Assumption P6 yields L moment conditions.

5.4. Model and moment selection in dynamic panel data models

We now show how to apply the MMSC and testing procedures of Section 2 to model (5.1) using the moment conditions of the previous subsection.

For a given restricted version of model (5.1), let θ denote the parameter vector that includes all parameters that enter the restricted model, as in (5.7) or (5.9). Let p denote the dimension of θ. The largest θ can be is as in (5.6), which corresponds to the general case where the parameter vector takes different values at each time period. The set of possibly correct moment conditions for a given restricted version of model (5.1) is a specified subset of (5.14)–(5.23). Let r denote the total number of these moment conditions. Then, a pair of model and moment selection vectors (b, c) consists of a $p \times 1$ vector b and an $r \times 1$ vector c, both containing zeros and ones. Zeros in b indicate that the model does not depend on the corresponding parameters in θ and zeros in c indicate that the corresponding moment conditions are not employed when estimating the parameters in $\theta_{[b]}$.

The parameter space $\mathscr{BC}$ for (b, c) should incorporate a considerable amount of information in order to eliminate many combinations of b and c. First, for a given restricted model, most variables in the model will be known to enter the model. Hence, $\mathscr{BC}$ will only contain b vectors with ones corresponding to the coefficients on these variables. Second, for most variables, the type of the variable will be known or partly known, be it predetermined, strictly exogenous, correlated with η_i, and/or uncorrelated with η_i. Hence, $\mathscr{BC}$ will only contain c vectors with ones corresponding to the appropriate moment conditions.

Third, the moment conditions in (5.14)–(5.23) typically are included or not included for all relevant time periods, such as $t = 1, \ldots, T$, rather than time period by time period. The parameter space $\mathscr{BC}$ is defined accordingly. Fourth, moment conditions in (5.17) and (5.22) are not included at the same time and those in (5.22) are included only if those in (5.20) are included.

Lastly, any other information about the correct parameter and moment vectors also should be used. Such information helps to reduce the parameter space and ease the selection problem.

For any $(b, c) \in \mathscr{BC}$, we evaluate the moment conditions selected by c at the parameters selected by b. Specifically, we substitute the following expression in each of the selected moment conditions in place of u_{it}:

$$y_{it} - w_{it}'(\delta + \delta_{(t)}), \tag{5.24}$$

where $\delta_{(1)} \equiv \mathbf{0}$ and each parameter in $(\delta', \delta_{(2)}', \ldots, \delta_{(T)}')'$ is set equal to zero if it is not included in θ or if the corresponding element in b is zero.

The weight matrix $W_n(b, c)$ for the GMM criterion function can be taken to equal $V_n^{-1}(b, c)$, where $V_n(b, c)$ is defined in footnote 2 with $\tilde{\theta}_n(b, c)$ equal to the GMM estimator of θ obtained by using the moments selected by c, the parameter space $\Theta_{[b]}$ for θ, and the weight matrix equal to the identity matrix.

Selection of the parameter vector and the moment conditions, including lag length, detection of structural breaks, exogeneity of regressors, etc., is conducted simultaneously. Given a model and moment selection estimator $(\hat{b}, \hat{c})$, the parameters $\theta_{[\hat{b}]}$ selected by $\hat{b}$ are estimated using the moment conditions selected by $\hat{c}$.

It remains to verify Assumption 1 of Section 2 for the dynamic panel data models considered in this section. This can be done for the case of observations that are identically distributed or non-identically distributed across individuals i.

For brevity, we just give sufficient conditions for the identically distributed case. Note that independence across individuals i has already been assumed. Assumption 1(a) holds for iid observations by the central limit theorem with $G^0(\theta)$ equal to the expectation of $G_n(\theta)$ provided $\mathrm{E}\|G_n(\theta)\|^2 < \infty \;\; \forall \theta \in \Theta$. All of the moments conditions in (5.14)–(5.23) just involve (at most) products of the underlying variables. In consequence, a sufficient condition for this moment condition is

$$\mathrm{E}\|(y_{i,1-L}, \ldots, y_{i,T}, x'_{it}, f'_i)'\|^4 < \infty. \tag{5.25}$$

The convergence part of Assumption 1(b) holds by a weak law of large numbers using the preceding moment conditions provided $\tilde{\theta}_n(b, c)$ converges in probability to some parameter $\theta^0(b, c)$ for each $(b, c) \in \mathscr{BC}$.[5] The matrix $W^0(b, c)$ equals $V^0(b, c)^{-1}$ in this case, where

$$\begin{aligned} V^0(b, c) = \mathrm{E}(&m_c(Z_i, \theta^0(b, c)) - \mathrm{E}m_c(Z_i, \theta^0(b, c))) \\ &(m_c(Z_i, \theta^0(b, c)) - \mathrm{E}m_c(Z_i, \theta^0(b, c)))'. \end{aligned} \tag{5.26}$$

The positive definiteness part of Assumption 1(b) holds if $V^0(b, c)$ is positive definite for all $(b, c) \in \mathscr{BC}$.

The convergence part of Assumption 1(c) holds using a Vapnik–Cervonenkis-type uniform weak law of large numbers for iid random variables under the moment conditions above using the linear or quadratic structure of the moment conditions, e.g., see Pollard (1984, Theorem II.24, Lemmas II.25 and II.27). The equality in Assumption 1(c) holds provided $\Theta_{[b]}$ is compact or $\Theta_{[b]}$ is of the form $\Theta_{[b]} = \{\theta \odot b\colon \theta \in R^p\}$ for all b such that $(b, c) \in \mathscr{BC}$ for some c, where '$\odot$' denotes element by element product.

6. Monte Carlo experiment

In this section, we conduct a Monte Carlo experiment to evaluate the performance of the MMSC and downward testing procedures. We consider MMSC–AIC, MMSC–BIC, MMSC–HQIC, and DT. We set $Q = 2.1$ in MMSC–HQIC.

The model we use is a restricted version of the general model in (5.1).

[5] For example, the latter holds if

$$Q_c(\theta^0(b,c)) < \inf_{\theta \in \Theta_{[b]}, \|\theta - \theta^0(b,c)\| > \varepsilon} Q_c(\theta) \quad \text{for all } \varepsilon > 0,$$

where

$$Q_c(\theta) = \mathrm{E}(m_c(Z_i, \theta)) - \mathrm{E}m_c(Z_i, \theta))'(m_c(Z_i, \theta) - \mathrm{E}m_c(Z_i, \theta)).$$

In turn, sufficient conditions for this are that $Q_c(\theta)$ is uniquely minimized over $\theta \in \Theta_{[b]}$ by $\theta^0(b,c)$ and $\Theta_{[b]}$ is compact.

6.1. The correct model

We consider a dynamic panel data model with lagged dependent variables and a covariate as regressors. We assume that the econometrician does not know the true lag length. We also assume that the econometrician does not know whether the covariate is correlated with the individual effect or whether the covariate is strictly exogenous with respect to the time-varying error component.

In particular, the correct model is

$$\begin{aligned} y_{it} &= \alpha_0 + \alpha_1 y_{i,t-1} + \beta x_{it} + u_{it}, \\ u_{it} &= \eta_i + v_{it} \quad \forall t = 1,\ldots,T \text{ and } i = 1,\ldots,N, \end{aligned} \tag{6.1}$$

where $\eta_i \sim \mathrm{N}(0, \sigma_\eta^2)$, $v_{it} \sim \mathrm{N}(0, \sigma_v^2)$, and $\mathrm{E}\eta_i v_{it} = 0$ for all t. The true lag length is one, i.e., $L_0 = 1$. The covariate x_{it} is predetermined, but not strictly exogenous with respect to the time-varying error v_{it}. It is correlated with the individual effect η_i for all t.

We take

$$(x_{i1},\ldots, x_{iT},\eta_i, v_{i1},\ldots, v_{iT})' \sim \mathrm{N}(\mathbf{0}, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \sigma_x^2 \mathbf{I}_T & \sigma_{x\eta}\mathbf{1}_T & \sigma_{xv}\Gamma \\ \sigma_{x\eta}\mathbf{1}_T' & \sigma_\eta^2 & \mathbf{0}_T' \\ \sigma_{xv}\Gamma' & \mathbf{0}_T & \sigma_v^2\mathbf{I}_T \end{pmatrix}. \tag{6.2}$$

Here, $\mathbf{I}_T$ denotes a $T \times T$ identity matrix, $\mathbf{1}_T$ denotes a $T \times 1$ vector of ones, $\mathbf{0}_T$ denotes a $T \times 1$ vector of zeros, Γ is a $T \times T$ matrix whose jkth element is one when $k = j - 1$ for $j = 2,\ldots,T$ and zero otherwise, $\sigma_{x\eta} = \mathrm{E}x_{it}\eta_i \neq 0$, and $\sigma_{xv} = \mathrm{E}x_{it}v_{it-1} \neq 0$. As specified, (i) x_{it} is uncorrelated with x_{is} for $t \neq s$ and has a constant variance σ_x^2, (ii) v_{it} is serially uncorrelated and uncorrelated with η_i and both error components have constant variances of σ_v^2 and σ_η^2, respectively, and (iii) x_{it} is correlated with the individual effect and is predetermined (because $\mathrm{E}x_{it}v_{is} = 0$ for $s = t + 1,\ldots,T$), but not strictly exogenous (because $\mathrm{E}x_{it}v_{it-1} = \sigma_{xv} \neq 0$ and $\mathrm{E}x_{it}v_{is} = 0$ for $s \neq t - 1$).

The L initial observations are specified by

$$\begin{aligned} y_{i,s} &= \alpha_0 + \alpha_1 y_{i,s-1} + \beta x_{is} + \eta_i + v_{is}, \quad s = 2 - L,\ldots, 0, \\ y_{i,1-L} &= \kappa + \frac{\beta\sigma_{x\eta} + \sigma_\eta^2}{\sigma_\eta^2(1-\alpha_1)}(\phi\eta_i + v_{i,1-L}), \end{aligned} \tag{6.3}$$

where $v_{i,1-L}, v_{i,s} \sim \mathrm{N}(0, \sigma_v^2)$, $\eta_i \sim \mathrm{N}(0, \sigma_\eta^2)$, $\phi = 1$, and $\kappa = \alpha_0/(1 - \alpha_1)$. The parameter ϕ controls the correlation between the initial observations and the individual effect η_i. The choice $\phi = 1$ implies that the 'stationarity' assumption, i.e., Assumption P6, holds. The parameter κ controls the mean levels of the initial observations. It does not affect whether the 'stationarity' assumption holds or not. It is chosen so the means of the observations are stationary.

In specifying the correct model, we use parameter values that have the following features: (i) there is a noticeable difference in efficiency between the GMM estimator that uses the correct model and all correct moment conditions and the GMM estimator that uses the least parsimonious model and only those moment conditions that are known to be correct and (ii) there are noticeable biases in the GMM estimators that are based on models that exclude some parameters whose true values are non-zero and/or use incorrect moment conditions. For parameter values with these features, there are gains to be exploited by a good selection procedure and losses to be incurred by a poor selection procedure. The following parameter values exhibit the desired features:

$$(\alpha_0,\ \alpha_1,\ \beta) = (0.8,\ 0.85,\ 0.5) \quad \text{and}$$
$$(\sigma_{x\eta},\ \sigma_{xv},\ \sigma_\eta^2,\ \sigma_v^2,\ \sigma_x^2) = (-0.2,\ 0.5,\ 1,\ 1,\ 1). \tag{6.4}$$

We want to examine how the selection procedures' finite sample performances change across both N and T. In consequence, we conduct experiments with five different sample size configurations: $(T, N) = (3, 250), (3, 500), (3, 1000), (6, 250)$, and $(6, 500)$. We employ 1000 simulation repetitions for each sample.

To evaluate the robustness of our results to models that exhibit a high degree of persistence, we also report results from one experiment with $\alpha_1 = 0.95$. We consider the sample size configuration $(T, N) = (3, 500)$. The full parameter vectors in this case are:

$$(\alpha_0,\ \alpha_1,\ \beta) = (0.8,\ 0.95,\ 0.5) \quad \text{and}$$
$$(\sigma_{x\eta},\ \sigma_{xv},\ \sigma_\eta^2,\ \sigma_v^2,\ \sigma_x^2) = (-0.2,\ 0.5,\ 0.2,\ 0.2,\ 5). \tag{6.5}$$

This case has received attention in the literature. Ahn and Schmidt (1995) and Blundell and Bond (1995) have shown that when α_1 is close to one, moment conditions based on the first differences of y_{it} may not be very informative, whereas moment conditions based on the 'stationarity' assumption can be very informative.

6.2. The parameter space for model and moment selection vectors

We assume that the econometrician does not know the correct model. Instead, he considers GMM estimation of the following model:

$$y_{it} = \alpha_0 + \alpha_1 y_{i,t-1} + \alpha_2 y_{i,t-2} + \beta x_{it} + u_{it},$$
$$u_{it} = \eta_i + v_{it} \quad \forall t = 1,\dots,T \text{ and } i = 1,\dots,N. \tag{6.6}$$

The econometrician selects a lag length of 0, 1, or 2. For possible moment conditions, he considers the following four groups of assumptions:

Assumption G1. (a) $\mathrm{E}\eta_i = 0$, $\mathrm{E}v_{it} = 0$, $\mathrm{E}v_{it}\eta_i = 0\ \forall t = 1, 2, \dots, T$.
(b) $\mathrm{E}v_{is}v_{it} = 0\ \forall s, t = 1, 2, \dots, T$ with $s \neq t$.
(c) $\mathrm{E}v_{it}y_{i0} = \mathrm{E}v_{it}y_{i,-1} = 0\ \forall t = 1, 2, \dots, T$.
(d) $\mathrm{Var}(v_{it}) = \sigma_i^2$ for some $\sigma_i^2 > 0\ \forall t = 1, 2, \dots, T$.
(e) $\mathrm{E}v_{it}(x_{i1}, \dots, x_{it}) = 0\ \forall t = 1, 2, \dots, T$.

Assumption G2. $\mathrm{E}v_{it}(x_{i,t+1}, \dots, x_{iT}) = 0\ \forall t = 1, 2, \dots, T-1$.

Assumption G3. $\mathrm{E}\eta_i x_{it} = 0\ \forall t = 1, 2, \dots, T$.

Assumption G4. $\mathrm{E}\eta_i y_{i1} = \mathrm{E}\eta_i y_{i0} = \mathrm{E}\eta_i y_{i,-1}$.

Assumption G1 imposes the standard error-component structure, constant variance for v_{it}, and predeterminedness for x_{it}. Assumption G2 further imposes strict exogeneity for x_{it}. Assumption G3 assumes x_{it} is uncorrelated with η_i. Assumption G4 is the 'stationarity assumption'.

Under the correct model, Assumptions G1 and G4 hold, but Assumptions G2 and G3 do not hold. We assume that the econometrician only knows that Assumption G1 holds. The econometrician determines the validity of Assumptions G2–G4 by using a selection procedure. For computational reasons in the Monte Carlo experiments, we only consider *linear* moment conditions. These conditions are the following:

Moment Conditions 1. (a) $\mathrm{E}(u_{i1}, \dots, u_{iT}) = 0$.
(b) $\mathrm{E}(y_{i,1-L}, \dots, y_{i,t-2})\Delta u_{it} = 0\ \forall t = 2, \dots, T$.
(c) $\mathrm{E}(y_{i,t-1}\Delta u_{it} - y_{it}\Delta u_{i,t+1}) = 0\ \forall t = 2, \dots, T-1$.
(d) $\mathrm{E}(x_{i1}, \dots, x_{i,t-1})\Delta u_{it} = 0\ \forall t = 2, \dots, T$.

Moment Conditions 2. $\mathrm{E}(x_{it}, \dots, x_{iT})\Delta u_{it} = 0\ \forall t = 2, \dots, T$.

Moment Conditions 3. $\mathrm{E}(u_{it} + \cdots + u_{iT})x_{it} = 0\ \forall t = 1, \dots, T$.

Moment Conditions 4. $\mathrm{E}(u_{i2} + \cdots + u_{iT})\Delta y_{i,t} = 0\ \forall t = 2 - L, \dots, 1$.

Moment Conditions j are implied by Assumptions G1 and Gj for $j = 1, \dots, 4$.

For the above model and moment selection problem, the largest parameter vector that the econometrician considers is

$$\theta = (\alpha_0, \alpha_1, \alpha_2, \beta)'. \tag{6.7}$$

We assume the econometrician always includes an intercept in the model, selects 0, 1, or 2 lags, and selects to include or exclude the covariate x_{it}. This yields six selection vectors b. The largest collection of moment conditions the econometrician considers includes all of the Moment Conditions 1–4. We assume that the econometrician knows that Moment Conditions 1 are correct and selects either all or none of the moment conditions in each group of Moment Conditions 2–4. This yields eight selection vectors c. Thus, the parameter space $\mathscr{BC}$ contains forty-eight (b, c) pairs. Each pair is a combination of one of the following six model selection vectors and eight moment selection vectors:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}. \tag{6.8}$$

The correct model selection vector is $b^0 = (1, 1, 0, 1)'$ and the correct moment selection vector is $c^0 = (1, 0, 0, 1)'$, which selects the Moment Conditions 1 and 4.

6.3. Measures of performance

We report two sets of results that measure the performances of the MMSC and DT procedures. First, for each selection procedure, we calculate the probabilities that the procedure

1. selects (b^0, c^0);
2. selects 'Other Consistent (b, c)', i.e., $(b, c) \in \mathscr{BC}$ such that $b \geqslant b^0$, $c \leqslant c^0$, and $(b, c) \neq (b^0, c^0)$; and
3. selects 'Inconsistent (b, c)', i.e., $(b, c) \in \mathscr{BC}$ such that $b < b^0$ or $c > c^0$.

In the first case, the correct model and all correct moment conditions are selected and consistent parameter estimators are obtained. This is the ideal situation. In the second case, although (b^0, c^0) is not selected, the model and moment conditions selected lead to consistent GMM estimators. In the third case, the model and moment conditions selected lead to GMM parameter estimators that are inconsistent. A selection procedure with a high probability of selecting (b^0, c^0), coupled with a low probability of selecting 'Inconsistent (b, c)', leads to an efficient GMM estimator. A selection procedure with a moderate to high probability of selecting 'Inconsistent (b, c)' leads to a GMM estimator with poor finite sample properties due to the biases resulting from employing too parsimonious a model and/or incorrect moment conditions.

Second, we report the biases, standard errors, and root mean-squared errors (RMSEs) of the post-selection GMM estimators for each selection procedure. We also report the rejection rates of the 5% t-tests based on the post-selection

GMM estimators. (When a parameter is excluded from the selected model, its estimated value is set equal to zero when computing the t-statistic.) Each t-test tests the null hypothesis that a parameter equals a value that is the true value and, hence, the null is true. The critical values for the t tests are the 5% critical values from a standard normal distribution.

In order to assess the performance of the post-selection GMM estimators, we also report biases, standard errors, etc. for four benchmark GMM estimators that are not post-selection estimators. The first such estimator is the GMM estimator based on the correct model and moment selection vector (b^0, c^0). This estimator is infeasible, but is used as a benchmark for good performance. The second estimator is the GMM estimator based on the least restrictive specification: $(b_{lr}, c_{lr}) = (\mathbf{1}_p, (1, 0, 0, 0)')$. The third estimator is the GMM estimator based on $(b, c) = (\mathbf{1}_p, \mathbf{1}_r)$, i.e., the whole parameter vector and all of the moment conditions. The fourth estimator is the GMM estimator based on the most restrictive specification: $(b_{mr}, c_{mr}) = ((1, 0, 0, 0)', \mathbf{1}_r)$. The second through fourth estimators are feasible estimators. Given the correct model, the second leads to consistent GMM estimators and the econometrician knows this (given the assumptions). The third and fourth estimator do not lead to consistent GMM estimators, although the econometrician does not know this given the assumptions.

We refer to the post-selection estimators of θ based on MMSC–AIC, MMSC–BIC, MMSC–HQIC, and DT as GMM$(b_{\mathrm{AIC}}, c_{\mathrm{AIC}})$, GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$, GMM$(b_{\mathrm{HQIC}}, c_{\mathrm{HQIC}})$, and GMM$(b_{\mathrm{DT}}, c_{\mathrm{DT}})$ respectively. We refer to the four benchmark GMM estimators as GMM$(b^0\ c^0)$, GMM(b_{lr}, c_{lr}), GMM$(\mathbf{1}_p, \mathbf{1}_r)$, and GMM$(b_{mr}, c_{mr})$.

6.4. Monte Carlo results

Now we present the Monte Carlo results for the selection probabilities and post-selection estimators and tests. The results for the post-selection estimators and tests ultimately are of greatest interest. But, the results for the selection probabilities help explain the pattern of results obtained for the post-selection estimators and tests.

6.4.1. Selection probabilities

Table 1 reports the selection probabilities for MMSC–AIC, MMSC–BIC, MMSC–HQIC, and DT for six different sample size/parameter combinations. The first three combinations in Part A of the Table are for $\alpha_1 = 0.85$ and (T, N) equal to (3, 250), (3, 500), and (3, 1000). The effect of increasing N is quite dramatic. For MMSC–BIC, the probability of selecting (b^0, c^0) increases from 0.482 to 0.852 to 0.990; while the probability of selecting 'Inconsistent (b, c)' declines from 0.487 to 0.116 to 0.000. For MMSC–HQIC, the corresponding changes are from 0.663 to 0.855 to 0.918 and from 0.214 to 0.028 to 0.000

Table 1
Selection probabilities

	MMSC–AIC	MMSC–BIC	MMSC–HQIC[b]	DT
(A) $\alpha_1 = 0.85$[a]				
Sample size: $T = 3$, $N = 250$				
(b^0, c^0)	0.607	0.482	0.663	0.559
Other consistent (b,c)[c]	0.328	0.031	0.123	0.046
Inconsistent (b,c)[d]	0.065	0.487	0.214	0.395
Sample size: $T = 3$, $N = 500$				
(b^0, c^0)	0.664	0.852	0.855	0.915
Other consistent (b,c)	0.333	0.032	0.117	0.034
Inconsistent (b,c)	0.003	0.116	0.028	0.051
Sample size: $T = 3$, $N = 1000$				
(b^0, c^0)	0.658	0.990	0.918	0.955
Other consistent (b,c)	0.342	0.010	0.082	0.045
Inconsistent (b,c)	0.000	0.000	0.000	0.000
Sample size: $T = 6$, $N = 250$				
(b^0, c^0)	0.536	0.637	0.661	0.704
Other consistent (b,c)	0.458	0.115	0.283	0.250
Inconsistent (b,c)	0.006	0.248	0.056	0.046
Sample size: $T = 6$, $N = 500$				
(b^0, c^0)	0.622	0.928	0.850	0.859
Other consistent (b,c)	0.378	0.063	0.150	0.141
Inconsistent (b,c)	0.000	0.009	0.000	0.000
(B) $\alpha_1 = 0.95$[e]				
Sample size: $T = 3$, $N = 500$				
(b^0, c^0)	0.566	0.918	0.831	0.901
Other consistent (b,c)	0.428	0.033	0.156	0.079
Inconsistent (b,c)	0.006	0.049	0.013	0.020

[a] The true parameter values in Part A of the table are $(\alpha_0, \alpha_1, \alpha_2, \beta) = (0.8, 0.85, 0, 0.5)$ and $(\sigma_{x\eta}, \sigma_{xv}, \sigma_\eta^2, \sigma_v^2, \sigma_x^2) = (-0.2, 0.5, 1, 1, 1)$.

[b] $Q = 2.1$ in MMSC–HQIC.

[c] 'Other consistent (b,c)' refers to model and moment selection vectors other than (b^0, c^0) that yield GMM estimators that are consistent.

[d] 'Inconsistent (b,c)' refers to model and moment selection vectors that yield GMM estimators that are inconsistent.

[e] The true parameter values in Part B of the table are $(\alpha_0, \alpha_1, \alpha_2, \beta) = (0.8, 0.95, 0, 0.5)$ and $(\sigma_{x\eta}, \sigma_{xv}, \sigma_\eta^2, \sigma_v^2, \sigma_x^2) = (-0.2, 0.5, 0.2, 0.2, 5)$.

respectively. For DT, the corresponding changes are from 0.559 to 0.915 to 0.955 and from 0.395 to 0.051 to 0.000 respectively.

The selection probabilities of MMSC–AIC are much less sensitive to the sample size N than are those of the other three procedures. As the sample size N increases from 250 to 500 to 1000, the probability of selecting (b^0, c^0) by MMSC–AIC changes from 0.607 to 0.664 to 0.658 and the probability of selecting 'Inconsistent (b, c)' decreases from 0.065 to 0.003 to 0.000. The fact that the probability of selecting (b^0, c^0) does not increase toward one as N increases reflects the inconsistency of the MMSC–AIC procedure. For the smallest sample size, MMSC–AIC is the best of the three procedures. But, for larger samples sizes, it does not perform as well as the other two MMSC.

Next, we consider the cases where (T, N) equals (6, 250) and (6, 500). The effect of the increase in sample size N is quite similar to the case where $T = 3$. There is a dramatic improvement for MMSC–BIC, MMSC–HQIC, and DT, but only a modest improvement for MMSC–AIC.

The effect of fixing N at 250 or 500 and increasing T from 3 to 6 is quite similar to that of fixing T and increasing N. Specifically, the performances of MMSC–BIC, MMSC–HQIC, and DT improve dramatically, while that of MMSC–AIC changes relatively little.

The effect on the selection probabilities of increasing α_1 from 0.85 to 0.95 can be seen by comparing the results of Part B of Table 1 with those of Part A for $(T, N) = (3, 500)$. We find that MMSC–BIC improves somewhat, while MMSC–AIC, MMSC–HQIC, and DT deteriorate somewhat.

Overall, we find that MMSC–AIC works best for the smallest sample size $(T, N) = (3, 250)$, whereas MMSC–BIC, MMSC–HQIC, and DT work best for all other sample sizes. MMSC–BIC performs very well for the largest sample sizes. MMSC–BIC and DT appear to perform best in an all-around sense.

6.4.2. Post-selection estimation and testing

Tables 2–4 report biases, standard errors, etc. for the eight GMM estimators discussed above for the cases where $\alpha_1 = 0.85$ and (T, N) equals (3, 250), (3, 500), and (3, 1000) respectively. In each table, results for the four benchmark GMM estimators are listed on the left-hand side and those for the four post-selection GMM estimators are listed on the right-hand side.

In Tables 2–4, the benchmark estimators exhibit the following patterns. GMM(b^0, c^0) sets the standard for good performance. The consistent and feasible estimator GMM(b_{lr}, c_{lr}) has somewhat larger biases and much larger standard deviations and RMSEs than GMM(b^0, c^0) for α_0, α_1, and α_2. For example, for α_0 and α_1, its RMSEs are two to four times those of GMM(b^0, c^0). For β, its biases, standard deviations, and RMSEs are only marginally larger than those of GMM(b^0, c^0). Thus, there is considerable scope for the post-selection estimators to outperform GMM(b_{lr}, c_{lr}) in terms of RMSE for α_0, α_1, and α_2, but not for β. The rejection rates of the 5% tests for GMM(b_{lr}, c_{lr}) are noticeably higher than those for GMM(b^0, c^0) (and greater than 5%) when $N = 250$, but not for $N = 500$ or 1000.

The two inconsistent estimators GMM$(\mathbf{1}_p, \mathbf{1}_r)$ and GMM(b_{mr}, c_{mr}) perform very poorly. They have very large biases, standard errors, and RMSEs. Their rejection rates exceed the nominal 5% rate by a very large margin. These results indicate that the cost of using the wrong model and/or moment conditions in the

Table 2
Biases, standard deviations, and RMSEs of GMM estimators and rejection rates of 5% tests: $T = 3$, $N = 250$, $\alpha_1 = 0.85$[a,b]

	Bias	SD	RMSE	Rej. rate[c]	Bias	SD	RMSE	Rej. rate[c]
	GMM(b^0, c^0)[d]				GMM$(b_{\mathrm{AIC}}, c_{\mathrm{AIC}})$			
α_0	0.042	0.236	0.239	0.062	0.086	0.522	0.529	0.084
α_1	− 0.008	0.041	0.042	0.083	− 0.022	0.112	0.114	0.159
α_2	—	—	—	—	0.007	0.048	0.048	0.070
β	− 0.008	0.065	0.065	0.060	− 0.012	0.072	0.073	0.075
	GMM(b_{lr}, c_{lr})				GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$			
α_0	0.187	0.505	0.539	0.088	0.099	0.568	0.577	0.099
α_1	− 0.062	0.124	0.139	0.138	− 0.016	0.112	0.113	0.138
α_2	0.028	0.062	0.068	0.098	− 0.002	0.033	0.034	0.037
β	− 0.009	0.066	0.067	0.058	− 0.048	0.090	0.102	0.296
	GMM$(\mathbf{1}_b, \mathbf{1}_r)$				GMM$(b_{\mathrm{HQIC}}, c_{\mathrm{HQIC}})$[e]			
α_0	0.471	0.385	0.608	0.637	0.064	0.421	0.426	0.089
α_1	− 0.153	0.154	0.217	0.655	− 0.014	0.091	0.092	0.147
α_2	0.068	0.105	0.125	0.506	0.002	0.039	0.039	0.056
β	− 0.193	0.078	0.208	0.907	− 0.022	0.080	0.083	0.153
	GMM(b_{mr}, c_{mr})				GMM$(b_{\mathrm{DT}}, c_{\mathrm{DT}})$			
α_0	4.566	0.584	4.604	1.000	0.093	0.491	0.500	0.096
α_1	− 0.850	—	0.850	—	− 0.016	0.095	0.097	0.129
α_2	0.000	—	0.000	—	0.000	0.031	0.031	0.029
β	− 0.500	—	0.500	—	− 0.041	0.088	0.097	0.248

[a] The true parameter values are $(\alpha_0, \alpha_1, \alpha_2, \beta) = (0.8, 0.85, 0, 0.5)$ and $(\sigma_{x\eta}, \sigma_{xv}, \sigma_\eta^2, \sigma_v^2, \sigma_x^2) = (-0.2, 0.5, 1, 1, 1)$.

[b] The results are based on 1000 Monte Carlo repetitions.

[c] The rejection rate is the fraction of times the 5% t-test based on the given GMM estimator rejects the null hypothesis that the given parameter equals the true value.

[d] The GMM estimators are defined as in Section 5.4.3: GMM(b^0, c^0) – the GMM estimator based on the correct model and moment selection vectors; GMM(b_{lr}, c_{lr}) – the GMM estimator based on the least restrictive specification, where $b_{lr} = (1, 1, 1, 1)$ and $c_{lr} = (1, 0, 0, 0)$; GMM$(\mathbf{1}_b, \mathbf{1}_r)$ – the GMM estimator based on all of the parameters and moment conditions; GMM(b_{mr}, c_{mr}) – the GMM estimator based on the most restrictive specification, where $b_{mr} = (1, 0, 0, 0)$ and $c_{mr} = (1, 1, 1, 1)$; *GMM*$(b_{\mathrm{AIC}}, c_{\mathrm{AIC}})$, *GMM*$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$, and GMM$(b_{\mathrm{HQIC}}, c_{\mathrm{HQIC}})$ – the GMM estimators based on MMSC–AIC, MMSC–BIC, and MMSC–HQIC respectively.

[e] $Q = 2.1$ in MMSC–HQIC.

cases under consideration can be huge. There is ample room for the post-selection estimators to outperform GMM$(\mathbf{1}_p, \mathbf{1}_r)$ and GMM(b_{mr}, c_{mr}), but also the possibility that they will perform very poorly.

Table 3
Biases, standard deviations, and RMSEs of GMM estimators and rejection rates of 5% tests: $T = 3$, $N = 500$, $\alpha_1 = 0.85$[a]

	Bias	SD	RMSE	Rej. rate	Bias	SD	RMSE	Rej. rate
	GMM(b^0, c^0)				GMM(b_{AIC}, c_{AIC})			
α_0	0.027	0.152	0.155	0.057	0.025	0.241	0.242	0.077
α_1	– 0.006	0.026	0.026	0.060	– 0.006	0.063	0.063	0.123
α_2	—	—	—	—	0.001	0.033	0.033	0.063
β	– 0.005	0.045	0.046	0.065	– 0.005	0.046	0.046	0.059
	GMM(b_{lr}, c_{lr})				GMM(b_{BIC}, c_{BIC})			
α_0	0.068	0.329	0.336	0.053	0.029	0.172	0.174	0.078
α_1	– 0.024	0.082	0.085	0.070	– 0.005	0.038	0.038	0.103
α_2	0.011	0.044	0.045	0.065	– 0.001	0.017	0.017	0.023
β	– 0.005	0.047	0.047	0.056	– 0.013	0.055	0.057	0.141
	GMM($\mathbf{1}_b$, $\mathbf{1}_r$)				GMM(b_{HQIC}, c_{HQIC})			
α_0	0.469	0.263	0.538	0.838	0.021	0.190	0.191	0.076
α_1	– 0.141	0.106	0.176	0.749	– 0.004	0.050	0.050	0.111
α_2	0.054	0.071	0.089	0.494	– 0.001	0.027	0.027	0.055
β	– 0.201	0.056	0.208	0.991	– 0.007	0.047	0.048	0.073
	GMM(b_{mr}, c_{mr})				GMM(b_{DT}, c_{DT})			
α_0	4.519	0.381	4.535	1.000	0.030	0.166	0.168	0.068
α_1	– 0.850	—	0.850	—	– 0.006	0.035	0.035	0.080
α_2	0.000	—	0.000	—	0.000	0.014	0.014	0.012
β	– 0.500	—	0.500	—	– 0.008	0.050	0.050	0.097

[a] Footnotes 1–5 of Table 2 apply to this table as well.

The results given in Table 2 indicate that for $N = 250$ the post-selection estimators are roughly comparable in RMSE and rejection rate performance to GMM(b_{lr}, c_{lr}). Thus, they perform noticeably worse than GMM(b^0, c^0), but very much better than GMM($\mathbf{1}_p$, $\mathbf{1}_r$) and GMM(b_{mr}, c_{mr}). Given the rather small sample size, at least for panel data, these results are encouraging. Comparisons across the post-selection estimators exhibit mixed results for both RMSE and rejection rates. For β, GMM(b_{AIC}, c_{AIC}) is the best and GMM($b_{\mathrm{BIC}}, c_{\mathrm{BIC}}$) is the worst. For α_0 and α_1, GMM(b_{HQIC}, c_{HQIC}) is the best and GMM($b_{\mathrm{BIC}}, c_{\mathrm{BIC}}$) is the worst. For α_2, GMM($b_{\mathrm{BIC}}, c_{\mathrm{BIC}}$) and GMM($b_{\mathrm{DT}}, c_{\mathrm{DT}}$) are the best.

The results of Table 3 for $N = 500$ show that the post-selection estimators are much better than GMM(b_{lr}, c_{lr}) in terms of RMSE, although they are still worse

Table 4
Biases, standard deviations, and RMSEs of GMM estimators and rejection rates of 5% tests: $T = 3$, $N = 1000$, $\alpha_1 = 0.85$[a]

	Bias	SD	RMSE	Rej. rate	Bias	SD	RMSE	Rej. rate
	GMM(b^0, c^0)				GMM(b_{AIC}, c_{AIC})			
α_0	0.015	0.103	0.104	0.055	0.012	0.172	0.173	0.083
α_1	– 0.003	0.018	0.018	0.070	– 0.003	0.043	0.044	0.124
α_2	—	—	—	—	0.000	0.023	0.023	0.048
β	– 0.004	0.031	0.032	0.068	– 0.004	0.032	0.032	0.064
	GMM(b_{lr}, c_{lr})				GMM(b_{BIC}, c_{BIC})			
α_0	0.042	0.232	0.236	0.053	0.015	0.105	0.106	0.058
α_1	– 0.012	0.058	0.059	0.064	– 0.003	0.021	0.022	0.077
α_2	0.004	0.030	0.031	0.050	0.000	0.008	0.008	0.008
β	– 0.004	0.033	0.033	0.057	– 0.004	0.031	0.031	0.067
	GMM($\mathbf{1}_b$, $\mathbf{1}_r$)				GMM(b_{HQIC}, c_{HQIC})			
α_0	0.497	0.178	0.528	0.981	0.012	0.124	0.125	0.067
α_1	– 0.149	0.071	0.165	0.889	– 0.002	0.031	0.031	0.102
α_2	0.057	0.048	0.075	0.588	– 0.001	0.015	0.015	0.033
β	– 0.200	0.038	0.203	0.999	– 0.004	0.031	0.031	0.066
	GMM($b_{\mathrm{m}r}$, $c_{\mathrm{m}r}$)				GMM(b_{DT}, c_{DT})			
α_0	4.537	0.283	4.546	1.000	0.017	0.119	0.120	0.061
α_1	– 0.850	—	0.850	—	– 0.004	0.024	0.024	0.078
α_2	0.000	—	0.000	—	0.000	0.009	0.009	0.007
β	– 0.500	—	0.500	—	– 0.004	0.031	0.032	0.069

[a] Footnotes 1–5 of Table 2 apply to this table as well.

than GMM(b^0, c^0). They are somewhat worse than GMM(b^0, c^0) and GMM(b_{lr}, c_{lr}) in terms of rejection rates. The post-selection estimators are very much better than GMM($\mathbf{1}_p$,$\mathbf{1}_r$) and GMM(b_{mr}, c_{mr}) in terms of both RMSE and rejection rates. The ranking of the four post-selection estimators for RMSE and rejection rates is as follows. GMM($b_{\mathrm{DT}}, c_{\mathrm{DT}}$) is the best. GMM($b_{\mathrm{BIC}}, c_{\mathrm{BIC}}$) and GMM($b_{\mathrm{HQIC}}, c_{\mathrm{HQIC}}$) are slightly worse and GMM($b_{\mathrm{AIC}}, c_{\mathrm{AIC}}$) is the worst. The RMSE performances of GMM($b_{\mathrm{DT}}, c_{\mathrm{DT}}$) and GMM($b_{\mathrm{BIC}}, c_{\mathrm{BIC}}$) are much better than that of GMM($b_{\mathrm{AIC}}, c_{\mathrm{AIC}}$) and are not too far from that of GMM(b^0, c^0). These results reflect the selection probability results of Table 1. In sum, the results of Table 3 indicate that for a sample size of $(T, N) = (3, 500)$ post-selection estimators can outperform any of the feasible benchmark estimators with respect to RMSE.

Table 4 presents results for $(T, N) = (3,\ 1000)$. In this case, the performance of GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ is almost equivalent to that of GMM$(b^0,\ c^0)$ in terms of both RMSEs and rejection rates. Thus, GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ outperforms GMM(b_{lr}, c_{lr}) by a noticeable margin and totally dominates GMM$(\mathbf{1}_p,\ \mathbf{1}_r)$ and GMM(b_{mr}, c_{mr}). Its excellent performance could be anticipated from the results of Table 1, because it selects the correct model and moment conditions with very high probability. The performance of GMM$(b_{\mathrm{DT}}, c_{\mathrm{DT}})$ is close behind that of GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$. GMM$(b_{\mathrm{HQIC}},\ c_{\mathrm{HQIC}})$ and GMM$(b_{\mathrm{AIC}}, c_{\mathrm{AIC}})$ perform better than GMM(b_{lr}, c_{lr}), but neither is as good as GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ or GMM$(b_{\mathrm{DT}}, c_{\mathrm{DT}})$. The ordering of the four post-selection estimators in Table 4 is clear: GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ is first, GMM$(b_{\mathrm{DT}}, c_{\mathrm{DT}})$ is second, GMM$(b_{\mathrm{HQIC}}, c_{\mathrm{HQIC}})$ is third, and GMM$(b_{\mathrm{AIC}}, c_{\mathrm{AIC}})$ is fourth.

For brevity, we do not present post-selection estimation results for the sample sizes $(T,\ N) = (6,\ 250)$ and (6, 500). The results for these cases are similar to those of Tables 3 and 4, respectively, for $(T,\ N) = (3{,}5\ 00)$ and (3, 1000), which have the same total number of observations.

Lastly, in Table 5 we report results for the second set of parameter values and sample size $(T, N) = (3,\ 500)$. In this case, α_1 is close to one, so the dependent variable y_{it} is highly persistent and the 'stationarity assumption' Assumption G4 is very informative. In consequence, GMM(b_{lr}, c_{lr}), which does not exploit Assumption G4, is much less efficient than GMM(b^0, c^0). Its RMSEs are from seven to twenty times as large as those of GMM(b^0, c^0).

In Table 5, all four post-selection estimators outperform GMM(b_{lr}, c_{lr}) in terms of both RMSE and rejection rates, but all are outperformed by GMM(b^0, c^0) in terms of RMSE. The best post-selection estimator is GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ in terms of RMSE and rejection rates. Next best are GMM$(b_{\mathrm{DT}}, c_{\mathrm{DT}})$ and GMM$(b_{\mathrm{HQIC}}, c_{\mathrm{HQIC}})$. The RMSEs of GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ are roughly half the size of those of GMM$(b_{\mathrm{AIC}}, c_{\mathrm{AIC}})$. In addition, GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ performs very well in terms of rejection rates with rates of 0.050, 0.055, and 0.064 for α_0, α_1, and β.

In summary, the results of Tables 2–5 indicate that the MMSC and DT procedures are effective in delivering improved estimator performance over the feasible alternative benchmark estimators provided the sample size $(T,\ N)$ is greater than (3, 250). The improvement of the consistent MMSC and DT procedures as the sample size increases is quite evident. With a sample size of $(T,\ N) = (3,\ 1000)$, the GMM$(b_{\mathrm{BIC}},\ c_{\mathrm{BIC}})$ estimator performs as well as the infeasible estimator that relies on knowing the correct model and moment conditions. The choice of the best MMSC is unclear for the smallest sample size $(T,\ N) = (3,\ 250)$, but for all larger sample sizes it is clearly seen to be GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$. The DT procedure is comparable to GMM$(b_{\mathrm{BIC}}, c_{\mathrm{BIC}})$ in an overall sense. It performs slightly better for the smaller sample sizes, but slightly worse for the larger ones.

Table 5
Biases, standard deviations, and RMSEs of GMM estimators and rejection rates of 5% t-tests: $T = 3$, $N = 500$, $\alpha_1 = 0.95$[a,b]

	Bias	SD	RMSE	Rej. rate	Bias	SD	RMSE	Rej. rate
	GMM(b^0, c^0)				GMM(b_{AIC}, c_{AIC})			
α_0	0.057	0.189	0.197	0.039	0.692	1.862	1.987	0.140
α_1	− 0.004	0.012	0.012	0.035	− 0.040	0.108	0.115	0.141
α_2	—	—	—	—	− 0.003	0.015	0.016	0.064
β	− 0.003	0.010	0.010	0.056	− 0.020	0.052	0.056	0.138
	GMM(b_{lr}, c_{lr})				GMM(b_{BIC}, c_{BIC})			
α_0	1.670	2.162	2.732	0.184	0.136	1.047	1.056	0.050
α_1	− 0.094	0.124	0.156	0.182	− 0.009	0.062	0.062	0.055
α_2	− 0.010	0.020	0.022	0.068	0.000	0.007	0.007	0.021
β	− 0.046	0.060	0.075	0.188	− 0.005	0.030	0.031	0.064
	GMM($\mathbf{1}_b$, $\mathbf{1}_r$)				GMM(b_{HQIC}, c_{HQIC})			
α_0	0.286	0.526	0.599	0.525	0.295	1.326	1.359	0.078
α_1	− 0.052	0.030	0.060	0.832	− 0.018	0.078	0.079	0.081
α_2	0.034	0.019	0.039	0.679	− 0.001	0.011	0.011	0.041
β	− 0.045	0.012	0.047	0.995	− 0.009	0.038	0.039	0.086
	GMM(b_{mr}, c_{mr})				GMM(b_{DT}, c_{DT})			
α_0	15.211	0.427	15.217	1.000	0.277	1.323	1.352	0.074
α_1	− 0.950	—	0.950	—	− 0.016	0.076	0.077	0.074
α_2	0.000	—	0.000	—	− 0.001	0.010	0.010	0.024
β	− 0.500	—	0.500	—	− 0.009	0.037	0.038	0.090

[a] The true parameter values are $(\alpha_0, \alpha_1, \alpha_2, \beta) = (0.8, 0.95, 0, 0.5)$ and $(\sigma_{x\eta}, \sigma_{xv}, \sigma^2_\eta, \sigma^2_v, \sigma^2_x) = (-0.2, 0.5, 0.2, 0.2, 0.5)$.

[b] Footnotes 2–5 of Table 2 apply to this table as well.

7. Conclusions

This paper extends the standard GMM framework to the case where there is imperfect knowledge about the correct model and moment conditions. We introduce a class of model and moment selection criteria (MMSC) and downward testing procedures that consistently select the correct model and all of the correct moment conditions, but no others. The MMSC are based on a trade-off between the magnitude of the J statistic and the numbers of parameters and moment conditions employed. The trade-off is analogous to that made by model selection criteria in likelihood scenarios.

The paper applies the MMSC and testing procedures to GMM estimation of dynamic panel data models. In such models, different GMM estimators are based on different sets of assumptions concerning the covariances between different components of the model, such as error components, regressors, and initial conditions. The selection procedures can be used to help determine which of the covariance restrictions are correct. The selection procedures also can be used to help specify the model. For example, they can be used to select the lag length, detect structural breaks in the parameters, or determine which regressors to include.

Lastly, we conduct a Monte Carlo experiment to evaluate the finite sample performance of the selection procedures. We consider a dynamic panel data problem. We compute the probabilities that several MMSC and downward testing procedures select the correct model and moment conditions, as well as various combinations of incorrect model and moment conditions. We analyze the performance of post-selection GMM estimators in terms of their biases, standard deviations, root mean-squared errors, and t-test rejection rates. The MMSC–BIC and downward testing procedures are found to work quite well in a variety of contexts.

Appendix A. Proofs

Proof of Theorem 1. The proof is quite similar to that of Theorem 1 of Andrews (1999). First, we establish Theorem 1(a). For any $(b,c)\in\mathscr{BC}$ with $(b,c)\notin\mathscr{BCZ}^0$, we have

$$J_n(b,\,c)/n \xrightarrow{\text{p}} \inf_{\theta_{[b]}\in\Theta_{[b]}} G_c^0(\theta_{[b]})'W^0(b,\,c)G_c^0(\theta_{[b]}) > 0 \quad \text{under } P^0, \tag{A.1}$$

where the convergence holds by Assumption 1(c) and the inequality holds because (i) $G_c^0(\theta_{[b]}) \neq \mathbf{0}\ \forall\theta_{[b]}\in\Theta_{[b]}$ by the supposition that $(b,\,c)\notin\mathscr{BCZ}^0$ and (ii) $W^0(b,\,c)$ is positive definite by Assumption 1(b). Eq. (A.1) and Assumption MMSC(b) yield: For any $(b,\,c)\in\mathscr{BC}$ with $(b,\,c)\notin\mathscr{BCZ}^0$,

$$\begin{aligned}\text{MMSC}_n(b,c)/n &= J_n(b,c)/n - h(|c|-|b|)\kappa_n/n\\ &\xrightarrow{\text{p}} \inf_{\theta_{[b]}\in\Theta_{[b]}} G_c^0(\theta_{[b]})'W^0(b,c)G_c^0(\theta_{[b]}) > 0 \quad \text{under } P^0.\end{aligned} \tag{A.2}$$

For any $(b,c)\in\mathscr{BCZ}^0$, we have

$$J_n(b,c) = \text{O}_\text{p}(1) \quad \text{under } P^0, \tag{A.3}$$

using Assumptions 1(a) and (c) and the fact that $G_c^0(\theta_{[b]}) = \mathbf{0}$ for some $\theta_{[b]}\in\Theta_{[b]}$. Eq. (A.3) and Assumption MMSC(b) yield: For any $(b,c)\in\mathscr{BCZ}^0$,

$$\text{MMSC}_n(b,c)/n = \text{O}_\text{p}(1) - h(|c|-|b|)\kappa_n/n = \text{O}_\text{p}(1) \quad \text{under } P^0. \tag{A.4}$$

Eqs. (A.2) and (A.4) imply that $(\hat{b}_{\text{MMSC}},\,\hat{c}_{\text{MMSC}})\in\mathscr{BCZ}^0$ wp $\to$ 1.

Now, suppose (b_1, c_1), $(b_2, c_2) \in \mathscr{BCZ}^0$, $(b_1, c_1) \notin \mathscr{MBCZ}^0$, and $(b_2, c_2) \in \mathscr{MBCZ}^0$. Then, $|c_1| - |b_1| < |c_2| - |b_2|$ and by Assumption MMSC

$$(h(|c_1| - |b_1|) - h(|c_2| - |b_2|))\kappa_n \to -\infty\,. \tag{A.5}$$

Eqs. (A.3) and (A.5) imply that $\mathrm{MMSC}_n(b_1, c_1) > \mathrm{MMSC}_n(b_2, c_2)$ wp $\to 1$. Thus, $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}}) \in \mathscr{MBCZ}^0$ wp $\to 1$, as stated in Theorem 1(a).

Now, Assumption ID*bc* and $(b^0, c^0) \in \mathscr{BC}$ imply that $\mathscr{MBCZ}^0 = \{(b^0, c^0)\}$. Hence, coupled with Theorem 1(a), the former conditions imply that $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}}) = (b^0, c^0)$ wp $\to 1$. In addition, $(b^0, c^0) \in \mathscr{BC}$ is necessary for $(\hat{b}_{\mathrm{MMSC}}, \hat{c}_{\mathrm{MMSC}}) = (b^0, c^0)$. Hence, Theorem 1(b) holds.

Theorem 1(c) follows from Theorem 1(b). □

Proof of Theorem 2. First, we establish Theorem 2(a). For any $(b, c) \in \mathscr{BC}$ with $(b, c) \notin \mathscr{BCZ}^0$, we have

$$J_n(b, c)/\gamma_{n,|c|-|b|} \xrightarrow{p} \infty \quad \text{under } P^0 \tag{A.6}$$

by (A.1) and Assumption T. Thus, $\hat{k}_{\mathrm{DT}} \leqslant \#(\mathscr{MBCZ}^0)$ wp $\to 1$, where $\#(\mathscr{MBCZ}^0)$ denotes the (unique) number of over-identifying restrictions for the vector(s) in $\mathscr{MBCZ}^0$.

For any $(b, c) \in \mathscr{BCZ}^0$, (A.3) and Assumption T yield

$$J_n(b, c) < \gamma_{n,|c|-|b|} \text{ wp} \to 1 \quad \text{under } P^0. \tag{A.7}$$

In consequence, $\hat{k}_{\mathrm{DT}} = \#(\mathscr{MBCZ}^0)$ wp $\to 1$. This result and (A.6) imply that $(\hat{b}_{\mathrm{DT}}, \hat{c}_{\mathrm{DT}}) \in \mathscr{MBCZ}^0$ wp $\to 1$ and, hence, Theorem 2(a) holds.

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

References

Ahn, S.C., Schmidt, P., 1995. Efficient Estimation of Models for Dynamic Panel Data. Journal of Econometrics 68, 5–27.

Akaike, H., 1969. Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics 21, 243–247.

Akaike, H., 1977. On entropy maximization principle. In: Krishnaiah, P.R. (Ed.), Applications of Statistics. North-Holland, Amsterdam.

Amemiya, T., 1980. Selection of regressors. International Economic Review 21, 331–354.

Amemiya, T., MaCurdy, T.E., 1986. Instrumental-variable estimation of an error-components model. Econometrica 54, 869–881.

Anderson, T.W., Hsiao, C., 1982. Formulation and estimation of dynamic models using panel data. Journal of Econometrics 18, 47–82.

Andrews, D.W.K., 1992. Generic uniform convergence. Econometric Theory 8, 241–257.

Andrews, D.W.K., 1997. A stopping rule for the computation of generalized method of moments estimators. Econometrica 65, 913–931.

Andrews, D.W.K., 1999. Consistent moment selection procedures for generalized method of moments estimation. Econometrica 67, 543–564.

Andrews, D.W.K., Lu, B., 1999. Consistent model and moment selection criteria for GMM estimation with application to dynamic panel data models. Cowles Foundation Discussion Paper No. 1233, Yale University.

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies 58, 277–297.

Arellano, M., Bover, O., 1995. Another look at the instrumental variable estimation of error-components models. Journal of Econometrics 68, 29–51.

Balestra, P., Nerlove, M., 1966. Pooling cross section and time series data in the estimation of a dynamic model: the demand for natural gas. Econometrica 34, 585–612.

Bhargava, A., Sargan, J.D., 1983. Estimating dynamic random effects models from panel data covering short time periods. Econometrica 51, 1635–1659.

Blundell, R., Bond, S., 1995. Initial conditions and moment restrictions in dynamic panel data models. Working Paper No. W95/17, The Institute for Fiscal Studies, London.

Breusch, T.S., Mizon, G.E., Schmidt, P., 1989. Efficient estimation using panel data. Econometrica 57, 695–701.

Chamberlain, G., 1984. Panel data.. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. II. Elsevier, Amsterdam.

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

Gallant, A.R., Hsieh, D., Tauchen, G., 1997. Estimation of stochastic volatility models with diagnostics. Journal of Econometrics 81, 159–192.

Gallant, A.R., Tauchen, G., 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.

Hannan, E.J., 1982. Testing for autocorrelation and Akaike's criterion. In: Gani, J.M., Hannan, E.J. (Eds.), Essays in statistical science. Applied Probability Trust, Sheffield, pp. 403–412.

Hannan, E.J., Deistler, M., 1988. The Statistical Theory of Linear Systems. Wiley, New York.

Hannan, E.J., Quinn, B.G., 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.

Hausman, J.A., Taylor, W.E., 1981. Panel data and unobservable individual effects. Econometrica 49, 1377–1398.

Holtz-Eakin, D., Newey, W., Rosen, H.S., 1988. Estimating vector autoregressions with panel data. Econometrica 56, 1371–1395.

Kabaila, P., 1995. The effect of model selection on confidence regions and prediction regions. Econometric Theory 11, 537–549.

Keane, M.P., Runkle, D.E., 1992. On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous. Journal of Business and Economics Statistics 10, 1–9.

Kohn, R., 1983. Consistent estimation of minimal subset dimension. Econometrica 51, 367–376.

Kolaczyk, E.D., 1995. An information criterion for empirical likelihood with general estimating equations. Unpublished manuscript, Department of Statistics, University of Chicago.

Maddala, G.S., 1971. The use of variance components models in pooling cross section and time series data. Econometrica 39, 341–358.

Mundlak, Y., 1961. Empirical production function free of management bias. Journal of Farm Economics 43, 45–56.

Nishii, R., 1988. Maximum likelihood principle and model selection when the true model is unspecified. Journal of Multivariate Analysis 27, 392–403.

Pesaran, M.H., Smith, R.J., 1994. A generalized R^2 criterion for regression models estimated by the instrumental variables method. Econometrica 62, 705–710.

Phillips, P.C.B., Ploberger, W., 1996. An asymptotic theory of Bayesian inference for time series. Econometrica 64, 381–412.

Pollard, D., 1984. Convergence of Stochastic Processes. Springer, New York.

Pötscher, B.M., 1983. Order estimation in ARMA-models by Lagrangian multiplier tests. Annals of Statistics 11, 872–885.

Pötscher, B.M., 1989. Model selection under nonstationarity: autoregressive models and stochastic linear regression models. Annals of Statistics 17, 1257–1274.

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

Pötscher, B.M., Novák, A.J., 1994. The distribution of estimators after model selection: large and small sample results. Unpublished manuscript, Institute for Statistics, Operations Research, and Computer Science, University of Vienna.

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.

Sin, C.-Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207–225.

Smith, R.J., 1992. Non-nested tests for competing models estimated by generalized method of moments. Econometrica 60, 973–980.