A Consistent Method for the Selection of Relevant Instruments[1]

Alastair R. Hall

North Carolina State University[2]

Fernanda P. M. Peixe

University of Evora[3]

January 26, 2000

[1]This work was begun while Hall was a Senior Research Fellow and Peixe was a graduate student at the Department of Economics, University of Birmingham, UK, and this support is gratefully acknowledged. Peixe also gratefully acknowledges financial support from FCT under grant PRAXIS XXI/BD/13453/97. This work represents part of Peixe's PhD dissertation which is to be submitted to the University of Birmingham. We are grateful for the comments of Atsushi Inoue, Peter Schmidt and seminar participants at the Department of Economics, Michigan State University.

[2]Department of Economics, Box 8110, North Carolina State University, Raleigh NC 27695-8110, USA. Email: alastair_hall@ncsu.edu

[3]Departamento de Economia, Universidade de Evora, Apartado 74, 7001 Evora Codex, Portugal. Email: fmp@uevora.pt

Abstract

Generalized Method of Moments (GMM) is widely applied in econometrics. In most cases, there is a vast array of population moments upon which to base estimation and so the researcher must decide which moments to use. Andrews (1999) proposes a method for moment selection based on minimizing an information criterion which is the sum of the overidentifying restrictions test and a bonus term reflecting the number of overidentifying restrictions. In this paper, we consider the problem of moment selection in the case where generalized instrumental variables (GIV) estimation is used. In the literature on GIV, it is known that it is desirable to choose instruments on the basis of three attributes: orthogonality, relevance and uniqueness. It is shown that Andrews's (1999) method chooses instruments on the basis of the orthogonality property alone, and so leads to the inclusion of instruments which are irrelevant in the sense their inclusion has no impact on the asymptotic variance of the estimator. While this weakness is inconsequential asymptotically, it has an adverse effect on the finite sample properties. In this paper we propose a new method for selecting instruments on the basis of their relevance. This method is based on a canonical correlations information criterion which we believe to be new to the literature. It is shown that the method is consistent in the sense that it selects all relevant instruments from a candidate set of instruments which are orthogonal. It is also shown that the combination of Andrews's (1999) method and our own yields a consistent method for the selection of relevant, orthogonal instruments from a candidate set. Simulation evidence suggests the method works well.

1 Introduction

Generalized Method of Moments (GMM) (Hansen, 1982) has been widely applied in econometrics because it provides a computationally convenient method of estimation based on the type of information provided by economic models. This information takes the form of a population moment condition involving a data vector and the unknown parameter vector of interest. In many applications, the underlying economic model implies many such population moment conditions, and so researchers must decide which moments to use in the estimation. One response to this dilemma has been to estimate the model using a number of different choices of moment condition and then to use the overidentifying restrictions test to diagnoze which moments are compatible with the data. While frequently employed in practice, this strategy suffers from repeated testing problems which render inferences suspect. However, in a recent paper Andrews (1999) has provided a method of moment selection based on the overidentifying restrictions test which circumvents these problems. Andrews (1999) proposes an information criterion approach to the selection of moments in which the criterion consists of the overidentifying restrictions test and a bonus term dependent on the number of overidentifying restrictions.[1] For certain choices of bonus term, Andrews (1999) shows that minimization of this criterion over a particular set of moment conditions is a consistent method of moment selection in the sense that it selects with probability one the largest vector of moment conditions from this set which are compatible with the data.[2]

While Andrews's (1999) approach has circumvented the repeated testing problem, it possesses certain weaknesses as a method of moment selection. These are best understood by considering a leading case of GMM estimation, namely generalized instrumental variables estimators (GIV) (Hansen and Singleton, 1982) which is also the focus of our paper. Within this framework, the population moment condition takes the form $E[z_t \quad u_t(\theta_0)] = 0$

[1] It should be noted that Andrews (1999) also considers downward and upward testing strategies based on the overidentifying restrictions test.

[2] Strictly, this result is asymptotic and subject to certain important identification conditions.

where θ_0 is the parameter vector of interest. In most cases of interest, $u_t(\theta_0)$ is implied by the underlying economic model, and so the problem of moment selection reduces to one of choosing a vector of instruments from a set of candidates, $\mathcal{Z}$ say. Within the literature on GIV, it has been recognized that it is desirable for the selected instrument to possess three main properties: orthogonality, relevance and uniqueness.[3] Since Andrews's (1999) method is based on the overidentifying restriction test, it focuses purely on the orthogonality property. However, there is growing evidence that the other two attributes – and relevance in particular – are important determinants of both the asymptotic and finite sample properties of the estimator. For example, Staiger and Stock (1997) and Stock and Wright (1997) demonstrate that if the entire instrument vector is *ir*relevant, or nearly so, then standard asymptotic theory is inappropriate. Furthermore, even if the instrument vector is relevant, not all elements of the population moment condition may be informative about the parameter vector – or in the terminology of Breusch, Qian, Schmidt, and Wyhowski (1999), some elements of the population moment condition may be "redundant". Although the inclusion of redundant moment conditions does no harm asymptotically,[4] there is compelling evidence that it causes a deterioration in the finite sample properties of the estimator.[5] It therefore seems desirable that orthogonality should not be the only attribute upon which instruments are selected.

In this paper, we propose a method for selecting instruments based on their relevance. Although the analysis covers nonlinear models, it is most convenient to introduce the ideas within the context of the normal linear regression model and then present the extension to nonlinear models and non–normality afterwards. To start, it is necessary to formulate a

[3] *E.g.* See Hall, Rudebusch, and Wilcox (1996).

[4] The inclusion of additional valid population moment conditions can never increase the asymptotic variance of the estimator.

[5] *E.g.* See Hall and Peixe (1999) and the evidence reported below. It should also be noted that this point is generic to all GMM estimators; see Andersen and Sorensen (1996) for an illustration of this point for a case in which the population moment condition does not take the form defined above.

precise definition of the conditions under which an instrument is relevant or irrelevant. In Section 2, we show that the canonical correlations between the regressors and instruments provide a natural basis for a definition of relevance. We also present a definition of a nearly irrelevant instrument inspired by the work of Staiger and Stock (1997) and Stock and Wright (1997), and contrast our notion of irrelevance with the concept of redundancy introduced by Breusch, Qian, Schmidt, and Wyhowski (1999). It emerges that redundancy and irrelevance are closely related but have one important difference: redundancy is a conditional property whereas irrelevance is unconditional. More specifically, if we partition the instrument vector into $z_t = [z_{1t}', z_{2t}']'$ then whether or not z_{2t} is redundant depends on the particular choice for z_{1t}. In contrast, if z_{2t} is irrelevant then it is redundant for all possible choices of z_{1t}.

This connection between irrelevance and redundancy is exploited in the instrument selection method proposed below. In Section 3, we present a likelihood ratio statistic for testing the null hypothesis that z_{2t} is redundant conditional on z_{1t}, that is given the instrument vector takes the form $z_t = [z_{1t}', z_{2t}']'$. This statistic is a function of the sample canonical correlations, and is already familiar in statistics in other contexts. We show the test is consistent against the alternative that the instruments are not redundant. Given the relationship between irrelevance and redundancy described above, the irrelevance or relevance of z_{2t} can be diagnozed by testing whether z_{2t} is redundant for all possible choices of z_{1t}. However, such a strategy runs into the repeated testing problems mentioned above in the context of the overidentifying restrictions test. Therefore, in Section 4 we propose a method of instrument selection based on the minimization of an information criterion. This criterion consists of two functions: the first is the likelihood ratio statistic described above and so depends on the sample canonical correlations; the second is a penalty function which increases with the number of overidentifying restrictions. In view of this structure, we refer to the minimand as the *canonical correlations information criterion* or *CCIC* for short. It is shown that this method of instrument selection is consistent under certain

conditions on the penalty function. In this context, consistency means that the method excludes all irrelevant instruments with probability one in the limit. Section 5 shows how all the previous analysis can be extended to nonlinear models and non–normally distributed data.

In practice, it is desirable to choose instruments which are both orthogonal and relevant. Just as Andrews's (1999) method focuses on orthogonality and ignores relevance, our method focuses on relevance and ignores orthogonality. However, intuition suggests that a combination of the two methods should achieve the desired goal. Section 6 explores the properties of a strategy in which the two methods are applied sequentially. Given the nature of the problem, the most logical sequence is to apply Andrews's (1999) method first and then apply our method based on CCIC. It is shown that such a strategy includes all variables from the candidate set of instruments which are both orthogonal and relevant with probability one in the limit.

Section 7 concludes the paper and a mathematical appendix contains the proofs for all the main results and certain other technical details.

2 Irrelevant instruments: definition and consequences

As mentioned above, it is most convenient to introduce our method within the context of the univariate linear regression model. Accordingly, we consider the case in which

$$y_t = x_t^{'}\theta_0 + u_t, \qquad t = 1, 2, \ldots, T \tag{1}$$

and the $p \times 1$ parameter vector θ_0 is estimated via GMM based on the population moment condition

$$E[z_t u_t(\theta_0)] \; = \; 0 \tag{2}$$

where z_t is a $q \times 1$ vector of instruments and $u_t(\theta) = y_t - x_t^{'}\theta$. In this paper we are concerned with the case in which z_t is chosen from some candidate set of instruments, $\mathcal{Z}$. Therefore

we let Z_t denote a $q_{max} \times 1$ containing all members of $\mathcal{Z}$, and define $z_t = S_q Z_t$ for some $q \times q_{max}$ selection matrix S_q. Since the main focus of this paper is on the development of a new method for instrument selection, we adopt the following high level assumption about generation of $v_t^{'} = [x_t^{'}, Z_t^{'}, u_t]$.

Assumption 1 *(i)* $v_t \sim IN(0, \Sigma_v)$*; (ii)* Σ_v *is given by*

$$\Sigma_v = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} & \Sigma_{xu} \\ \Sigma_{zx} & \Sigma_{zz} & 0_{q_{max} \times 1} \\ \Sigma_{ux} & 0_{1 \times q_{max}} & \sigma^2 \end{bmatrix}$$

(iii) Σ_{xx}, Σ_{zz} *are respectively* $p \times p$, $q_{max} \times q_{max}$ *nonsingular matrices, and* σ^2 *is a positive scalar; (iv)* $rank\{\Sigma_{xz}\} = p$.

This rather restrictive distributional assumption is made purely to facilitate the exposition, and we discuss the ways in which it can be weakened in Section 5. However, there are two aspects of Assumption 1 which are not relaxed: (ii) implies the orthogonality of Z_t and u_t; and (iv) implies that θ_0 is identified by (2) with $z_t = S_q Z_t$ for at least one choice of S_q. Now let X, Z and y be respectively the $T \times p$, $T \times q_{max}$ and $T \times 1$ matrices with t^{th} rows $x_t^{'}$, $Z_t^{'}$ and y_t. Under Assumption 1, the optimal weighting matrix for the GMM estimation is $W_T = (S_q T^{-1} Z'Z S_q^{'})^{-1}$ and this leads to the IV estimator

$$\hat{\theta}_T = [X'Z S_q^{'}(S_q Z'Z S_q^{'})^{-1} S_q Z'X]^{-1} X'Z S_q^{'}(S_q Z'Z S_q^{'})^{-1} S_q Z'y \tag{3}$$

It can be shown that under Assumption 1

$$T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V_\theta) \tag{4}$$

where

$$V_\theta = \sigma^2 [\Sigma_{xz} S_q^{'}(S_q \Sigma_{zz} S_q^{'})^{-1} S_q \Sigma_{zx}]^{-1} \tag{5}$$

Equation (5) indicates that asymptotic variance of $\hat{\theta}_T$ depends in some way on the relationship between the instruments and regressors, but does not provide useful insights into

exactly what aspects of this relationship are important. However, more progress can be made by expressing the asymptotic variance in terms of the population canonical correlations between x_t and z_t. To introduce this alternative representation we need the following notation. Let $\{\rho_i; i = 1, 2, \dots p\}$ be the canonical correlations between x_t and z_t, and assume $\rho_i \geq \rho_{i+1}$. Let F and G be the $p \times p$ and $p \times q$ matrices whose i^{th} rows, $f_i^{'}$ and $g_i^{'}$ contain the weights in the linear combinations associated with the i^{th} canonical correlation, that is $Corr(f_i^{'} x_t, g_i^{'} z_t) = \rho_i$. Bowden and Turkington (1984)[p. 29-32] show that

$$V_\theta \;=\; \sigma^2 F R^{-2} F' \tag{6}$$

where $R = diag\,(\rho_1 \dots \rho_p)$. Equation (6) reveals that the asymptotic variance depends crucially on the population canonical correlations.

Equation (6) is the key to our definition of an irrelevant instrument. To present this definition, we introduce the following notation. Let $z_{t,j}$ be the j^{th} element of Z_t, and $S_{j,q}$ be a $q \times q_{max}$ selection matrix whose j^{th} column consists of zeros. Finally, let $\rho_i[a_t : b_t]$ to denote the i^{th} population canonical correlation between a_t and b_t.

Definition 1 *Irrelevant instrument*

$z_{t,j}$ is said to be irrelevant (for the estimation of θ_0) if

$$\rho_i \left[x_t : \{z_{t,j}, S_{j,q} Z_t\}\right] = \rho_i \left[x_t : S_{j,q} Z_t\right]$$

for all $i = 1, 2 \dots p$, all $q = p, p+1, \dots q_{max} - 1$ and all $S_{j,q}$.

This definition states that $z_{t,j}$ is irrelevant if its inclusion in the instrument vector, z_t, has no impact on the population canonical correlations *regardless* of the other variables included in z_t. We adopt the following definition of relevance.

Definition 2 *Relevant instrument*

$z_{t,j}$ is said to be relevant if

$$\rho_i \left[x_t : \{z_{t,j}, S_{j,q} Z_t\}\right] \;\geq\; \rho_i \left[x_t : S_{j,q} Z_t\right]$$

for all $i = 1, 2 \ldots p$, *all* $q = p, p+1, \ldots q_{max} - 1$ *and all* $S_{j,q}$, *and*

$$\rho_i \left[x_t : \{ z_{t,j}, S_{j,q} Z_t \} \right] \; > \; \rho_i \left[x_t : S_{j,q} Z_t \right]$$

for some i, *all* $q = p, p+1, \ldots q_{max} - 1$ *and all* $S_{j,q}$.

Therefore $z_{t,j}$ is considered relevant if its inclusion increases $\sum_{i=1}^{p} \rho_i$ for all possible choices for the remaining variables in the instrument vector, z_t. Notice this definition is a stronger statement than simply the converse of Definition 1. There is one other consequence of Definitions 1 and 2 which should be noted. If $Z_t(R)$ are relevant and $Z_t(I)$ are irrelevant then the correlation between $Z_t(R)$ and $Z_t(I)$ must be zero. Otherwise, $Z_t(I)$ would inherit part of the correlation between x_t and $Z_t(R)$ if the latter are omitted from the instrument vector.

It might be argued that this definition of irrelevance is too strong because in practice the inclusion of an additional instrument is likely to increase the canonical correlations by some amount even if this amount is only small. Such a scenario is considered by Staiger and Stock (1997) and Stock and Wright (1997). These authors develop a framework to analyze the consequences of nearly unidentified parameters for the limiting distribution of various estimators and their associated statistics. We follow their approach to introduce the following definition of a nearly irrelevant instrument.

Definition 3 *Nearly irrelevant instrument*

$z_{t,j}$ *is said to be nearly irrelevant instrument if*

$$\rho_i \left[x_t : \{ z_{t,j}, S_{j,q} Z_t \} \right] = \rho_i \left[x_t : S_{j,q} Z_t \right] \; + \; \eta_i T^{-1/2}$$

for all $i = 1, 2 \ldots p$, *all* $q = p, p+1, \ldots q_{max} - 1$, *all* $S_{j,q}$ *and* $\eta_i \neq 0$ *for at least one* i.

It is easily verified that the asymptotic variance of $\hat{\theta}_T$ is uneffected by the inclusion of nearly irrelevant instruments.

It is instructive to compare our definition of instrument irrelevance with the concept of moment redundancy recently introduced by Breusch, Qian, Schmidt, and Wyhowski (1999).

In general terms, a set of of moment conditions, $g_2(\theta_0) = 0$, is redundant for the estimation of θ_0 given another set of moment conditions, $g_1(\theta_0) = 0$, if estimation based on both sets of conditions together does not improve asymptotic efficiency relative to the estimation based on $g_1(\theta_0) = 0$ alone. This idea is quite general, and so to facilitate a comparison with our definition of irrelevance, we must first specialize the definition of redundancy to our setting. To this end, we partition the instrument vector in (2) into $z_t = (z_{1t}', z_{2t}')'$ where z_{it} is $q_i \times 1$ for $i = 1, 2$, and define $\Sigma_{ix} = E[z_{it}x_t']$, $\Sigma_{ij} = E[z_{it}z_{jt}']$ for $i, j = 1, 2$.

Definition 4 *Redundant moment condition*

The set of moment conditions $E[z_{2t}u_t(\theta_0)] = 0$ *is redundant for the estimation of* θ_0 *given* $E[z_{1t}u_t(\theta_0)] = 0$*:*[6]

$$\Sigma_{2x} = \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{1x} \tag{7}$$

The link between this concept and instrument relevance is expressed in the following theorem.

Theorem 1 *Let Assumption 1 hold. If* $E[z_{2t}u_t(\theta_0)] = 0$ *is redundant for the estimation of* θ_0 *given* $E[z_{1t}u_t(\theta_0)] = 0$ *then* $\rho_i[x_t : z_t] = \rho_i[x_t : z_{1t}]$ *for all* $i = 1, 2 \ldots p$.

Theorem 1 indicates the key difference between redundancy and irrelevance. If $E[z_{2t}u_t(\theta_0)] = 0$ is redundant given $E[z_{1t}u_t(\theta_0)] = 0$ then the canonical correlations between x_t and z_t are the same as those between x_t and z_{1t}. However, this statement is specific to the particular vector z_{1t}. In other words, redundancy is a conditional property. In contrast, if z_{2t} is irrelevant then it has no impact on the canonical correlations for any vector z_{1t}, and so irrelevance is an unconditional property. This difference becomes important below. In Section 4, we introduce a method for instrument selection and establish its consistency. The latter result must be premised on an identification condition. Such a condition is easily

[6] Breusch, Qian, Schmidt, and Wyhowski (1999)[Theorem 1] provide four equivalent conditions for redundancy. Here we have used their Condition (C).

stated in terms of relevant and irrelevant (or nearly irrelevant) instruments because of the unconditional nature of these definitions. It is far more problematic using a conditional property such as redundancy. We discuss this issue further in Section 7. From here on, for ease of exposition, we shall say that z_{2t} is redundant given z_{1t} if it satisfies Definition 4 rather than ascribing this property to the associated population moment conditions.

To conclude this section, we consider the consequences of including irrelevant instruments. From an asymptotic perspective, there is no cost because it can be shown that the inclusion of additional valid instruments can never increase the asymptotic variance of the estimator.[7] However, this reassuring conclusion does not apply in finite samples. To illustrate, we report results from a small simulation study. Artificial data were generated for the model specified by (1) and Assumption 1 with $p = 1$ and $q_{max} = 10$. The matrix Σ_{xz} is defined implicitly by generating x_t via

$$x_t \;=\; Z_t^{'}\pi \;+\; v_t$$

where π is a 10×1 vector whose first two elements are 0.5 and whose remaining elements are zero. This specification implies that $\{z_{t,1}, z_{t,2}\}$ are relevant and $\{z_{t,j}; j = 3, 4, \ldots 10\}$ are irrelevant by our definitions. Data for y_t are generated from (1) with $\theta_0 = 0$. We then consider the behaviour of the IV estimator in (3) with the 10 choices of instrument $z_t = z_{q,t}$ where $z_{1,t} = z_{t,1}$ and $z_{q,t}^{'} = (z_{q-1,t}^{'}, z_{t,q})'$ for $q = 2 \ldots 10$. Notice that with this construction $q = 1$ involves one of the relevant instruments, and the move from $q = 1$ to $q = 2$ introduces the second relevant instrument. However all subsequent increases in q introduce irrelevant instruments. On each replication, we compute t–statistic for $H_0 : \theta_0 = 0$, the empirical size of the test (at a 10% nominal significance level), the simulated bias and simulated mean squared error (MSE) of $\hat{\theta}_T$. The sample size is set to $T = 100$ observations and the calculations are based on 1000 replications.

[7] *E.g.* See White (1984).

The results are presented in Table 1. It can be seen that all the statistics appear reasonabley well behaved for $q = 1$ or 2 but that the inclusion of irrelevant instruments increases both the biases of $\hat{\theta}_T$ and the t–statistic, and also the degree of overrejection of the t test. Clearly, the biases and degree of overrejection increase with the number of irrelevant instruments, and to such an extent that the t–statistic has an empirical size of nearly twice its nominal value for $q = q_{max} = 10$. In contrast, the MSE drops rather dramatically with the inclusion of the second relevant instrument and then holds fairly constant, although it increases slowly with q once at least two irrelevant instruments are included.

This evidence suggests that it is beneficial for inference in finite samples to base estimation upon only those instruments which are relevant. In the next section, we present a statistic for testing whether z_{2t} is redundant conditional on z_{1t}. Section 4 uses this statistic to construct a method for the selection of all relevant instruments out of a candidate set.

3 Testing redundancy

In this section we present a statistic for testing whether a particular subset of the instrument vector is redundant. To facilitate this analysis, we now specify the following reduced form for x_t,

$$x_t = \Pi z_t + v_t \tag{8}$$

$$= \Pi_1 z_{1t} + \Pi_2 z_{2t} + v_t \tag{9}$$

where $\Pi = [\Pi_1, \Pi_2]$, Π_i is a $p \times q_i$ matrix, $q_1 + q_2 = q$ and $z_t' = [z_{1t}', z_{2t}']$ is partitioned conformabley. It is assumed that (8) is the correct specification for x_t and so the following condition holds.

Assumption 2 *Conditional on Z_t, $v_t \sim IN(0,\text{-})$.*

The covariance matrix - can be derived from Σ_v but its precise definition does not matter for our analysis. Below we derive a test for the redundancy of z_{2t}, and so to ensure that θ_0

is identified by (2) we impose the following condition.

Assumption 3 *rank*$\{E\left[z_{1t}x_t^{'}\right]\} = p$.

The log-likelihood function of a sample of T observations is given by

$$\begin{aligned} L\left(\text{-} \ ,\Pi\right) &= \left(-Tp/2\right)\ln\left(2\pi\right) - \left(T/2\right)\ln|\text{-}\ | \\ &\quad - \left(1/2\right)\sum_{t=1}^{T}\left[\left(x_t - \Pi z_t\right)\left(x_t - \Pi z_t\right)'\right] \end{aligned} \tag{10}$$

The null and alternative hypotheses of interest are respectively,

$$\begin{aligned} H_0 &: \ z_{2t} \text{ is redundant given } z_{1t} \\ H_1 &: \ z_{2t} \text{ is not redundant given } z_{1t} \end{aligned}$$

Within our framework, the statement that z_{2t} is redundant given z_{1t} is equivalent to the restriction $\Pi_2 = 0$, and so we can test H_0 versus H_1 using the likelihood ratio test for the hypothesis $\Pi_2 = 0$. This statistic is:[8]

$$LR_T = -T\sum_{i=1}^{p}\ln\left(1-\hat{\rho}_i^2\right) + T\sum_{i=1}^{p}\ln\left(1-r_i^2\right) \tag{11}$$

where the $\hat{\rho}_i$ are the sample canonical correlations between x_t and z_t, and r_i are the sample canonical correlations between x_t and z_{1t}. This statistic is well known in statistics but to our knowledge has not been proposed as a test for redundancy *per se*.[9] The following theorem establishes the test's asymptotic properties under H_0 and H_1.

Theorem 2 *Let Assumptions 1–3 hold. (i) If* H_0 *is satisfied then* $LR_T \xrightarrow{d} \chi^2_{pq_2}$*; (ii) If* H_1 *holds then* $T^{-1}LR_T \xrightarrow{p} k$ *for some positive constant* k.

Theorem 2(ii) implies that LR_T diverges a rate T under H_1 and hence that the test is consistent against this alternative.

[8] *E.g.* See Hamilton (1994)[p. 649].

[9] *E.g.* See Anderson (1984)[p.317].

This statistic can be used to test whether z_{2t} is redundant given z_{1t}. Suppose now that it is desired to determine whether z_{2t} is irrelevant. As mentioned above, this hypothesis is equivalent to the statement that z_{2t} is redundant given all possible choices of z_{1t}. Therefore, one way to assess the relevance or irrelevance of z_{2t} is to apply the test for redundancy repeatedly using different choices of z_{1t}. However, it is hard – if not impossible – to determine the overall significance level of such a testing strategy because of the dependencies between the statistics. A more attractive approach is to use an information criterion to determine which instruments are relevant. This is the topic of the next section.

4 Canonical correlations information criteria

Before we present the method for selecting relevant instruments, it is useful to impose the following structure on the set of candidate instruments.

Assumption 4 *(i)* $Z_t = [Z_t(R)', Z_t(I)']'$ *where* $Z_t(R)$ *is a* $q_R \times 1$ *vector of instruments which are relevant for the estimation of* θ_0 *and* $Z_t(I)$ *is* $q_I \times 1$ *vector of instruments which are irrelevant for the estimation of* θ_0*; (ii)* S^0 *is the* $q_R \times q_{max}$ *selection matrix which satisfies* $Z_t(R) = S^0 Z_t$.

Assumption 4(i) defines the decomposition of the candidate set into relevant and irrelevant instruments. Notice that Assumption 1(iv) and Definitions 1 and 2 imply $rank\{E[x_t Z_t(R)']\} = p$, and so θ_0 is identified by $E[Z_t(R)u_t(\theta_0)] = 0$. This condition also stipulates that $q_R \geq p$. However, there are no restrictions on q_I, which may be zero.

Since any instrument vector can be written as $S_q Z_t$, the problem of instrument selection can be viewed as a search for the appropriate selection matrix from the set,

$$\mathcal{S} = \{\, S_q \in \Re^{q \times q_{max}}; S_q \text{ is a selection matrix, } rank(S_q) = q;\ q = p, p+1, \ldots q_{max} \,\}$$

Notice that the selection matrix S^0 is contained in $\mathcal{S}$.

For a given choice of S_q, the canonical correlations information criterion (CCIC) is defined to be

$$CCIC(S_q) \;=\; \sum_{i=1}^{p} \ln\left(1 - r_i^2\right) \;+\; (q-p)\frac{f(T)}{T} \tag{12}$$

where $\{r_i; i = 1, 2 \ldots p\}$ are the sample canonical correlations between x_t and $S_q Z_t$, $(q-p)$ is the degree of overidentification and $f(T)$ is a function of the sample size. It is well known in the literature on information criteria that $f(T)$ must satisfy certain properties in order to establish consistency results. Guided by the earlier work, we impose the following condition.

Assumption 5 *$f(T) \to \infty$ as $T \to \infty$ and $f(T) = o(T)$.*

Inspection of $CCIC(S_q)$ reveals that its two components move in opposite directions in response to the inclusion of an additional instrument: the first term, $\sum_{i=1}^{p} \ln\left(1 - r_i^2\right)$, can never increase and the second term increases. Therefore, if the selection matrix is chosen to minimize $CCIC(S_q)$ then resulting instrument vector retains only those instruments whose inclusion reduces $\sum_{i=1}^{p} \ln\left(1 - r_i^2\right)$ sufficiently to offset the increase in the penalty function, $(q-p)f(T)/T$. Let the chosen selection matrix be $\tilde{S}$, that is

$$\tilde{S} \;=\; argmin_{S_q \in \mathcal{S}}\, CCIC(S_q) \tag{13}$$

The following theorem establishes the consistency of this method.

Theorem 3 *If Assumptions 1, 4 and 5 hold then $\tilde{S} \xrightarrow{p} S^0$.*

In other words, if the selection matrix is chosen to minimize $CCIC(S_q)$ then the chosen instrument vector contains only the relevant instruments with probability one as $T \to \infty$. The three most popular choices of $f(T)$ are 2, $ln(T)$ and $Qln(ln(T))$ (for $Q > 2$) which are associated respectively with the Akaike (1974), Schwarz (1978) and Hannan and Quinn (1979) information criteria. For simplicity, we refer to the CCIC with these three choices of $f(T)$ as CCAIC, CCBIC and CCHQIC respectively. Inspection reveals that the choices

of $f(T)$ for CCBIC and CCHQIC satisfy Assumption 5 but the choice for CCAIC does not. This leads to the following corollary of Theorem 3.

Corollary 1 *If Assumptions 1 and 3 hold then CCBIC and CCHQIC are consistent but CCAIC is inconsistent.*

The inconsistency of $CCAIC$ is one sided in the sense that it includes all relevant instruments with probability one in the limit but there is a non zero probability in the limit that irrelevant instruments are also included.

As remarked above in Section 2, the definition of irrelevance is rather strong because it implies such instruments make no contribution to the population canonical correlations regardless of the other instruments included. This motivated our definition of nearly irrelevant instruments, and we now consider the properties of CCIC in this case. So we replace Assumption 4 by the following condition.

Assumption 6 *(i)* $Z_t = [Z_t(R)', Z_t(NI)']'$ *where* $Z_t(R)$ *is a* $q_R \times 1$ *vector of instruments which are relevant for the estimation of* θ_0 *and* $Z_t(NI)$ *is* $q_{NI} \times 1$ *vector of instruments which are nearly irrelevant for the estimation of* θ_0*; (ii)* S^0 *is the* $q_R \times q_{max}$ *selection matrix which satisfies* $Z_t(R) = S^0 Z_t$.

The following theorem establishes the method is still consistent.

Theorem 4 *If Assumptions 1, 5 and 6 hold then* $\tilde{S} \xrightarrow{p} S^0$.

The definition of $\tilde{S}$ in (13) requires estimation with all possible selection matrices in $\mathcal{S}$ and this may be computationally burdensome if q_{max} is relatively large. However, inspection of the proof of Theorem 3 reveals that $CCIC$ increases with probability one asymptotically if a relevant instrument is removed but decreases with probability one asymptotically if an irrelevant is removed. This suggests the following simplified selection strategy.

Definition 5 *Simplified selection strategy*

The chosen instrument vector is $\hat{S}Z_t$ *where* $\hat{S}$ *is constructed as follows. Calculate the*

criteria with all instruments; let this value be $CCIC(I_{q_{max}})$. *For* $j = 1, 2, \ldots q_{max}$, *calculate the criteria with* $z_t = S_{j,q_{max}} Z_t$ *for* $j = 1, 2 \ldots q_{max}$ *where* $S_{j,q_{max}}$ *is the* $(q_{\max} - 1) \times q_{max}$ *constructed by deleting the* j^{th} *row of* $I_{q_{max}}$. *Let* $\{j_i;\, i = 1, 2 \ldots q\}$ *be values of* j *for which* $CCIC(S_{j,q_{max}}) > CCIC(I_{q_{max}})$. *Then* $\hat{S}$ *is the* $q \times q_{max}$ *matrix with* i^{th} *row equal to the* j_i^{th} *row of* $I_{q_{max}}$.

The following Corollary to Theorems 3 and 4 gives the properties of $\hat{S}$.

Corollary 2 *If the conditions of either Theorems 3 or 4 then* $\hat{S} \xrightarrow{p} S^0$.

To conclude this section, we discuss the results from a simulation study designed to investigate the finite sample properties of both the instrument selection strategies described above. The model is the same as in section 2 except we now set $q_{max} = 8$, and so the relevant instruments are $\{z_{t,1}, z_{t,2}\}$ and the other six, $\{z_{t,j};\, j = 3, 4 \ldots 8\}$, are irrelevant. Even with this simple design, there are total of 255 possible instrument vectors. We begin with the simplified selection strategy given in Definition 5. Table 2 reports the frequency with which each of the eight instruments is selected. It is clear that all three criteria tend to pick the two relevant instruments with probability very close to one if $T = 100$ and equal to one if $T = 500$. The main difference in the three criteria appears to be in the frequency with which they select irrelevant instruments. Clearly, CCAIC includes irrelevant instruments most frequently which is in line with Corollary 1. Of the two consistent methods, CCBIC includes irrelevant instruments less frequently as would be expected from the choices of $f(T)$ associated with the criteria. Table 2 also reports the mean and empirical size of the t-statistic for $\theta_0 = 0$ based on the IV estimator calculated with $\hat{S}Z_t$. For $T = 100$, CCBIC appears to lead to a t-statistic whose behaviour is most closely approximated by asymptotic theory and this reflects the value of excluding irrelevant instruments. By $T = 500$, the empirical size of all three versions of the t-statistic is very close to the nominal value of 0.1. This evidence suggests the simplified instrument selection strategy works well.

We now consider the behaviour of the original strategy defined by (13). In view of the large number of possible combinations of instruments, it was decided to limit the search by choosing the instruments in pairs. One pair consists of the two relevant instruments, that is (z_1, z_2), and the remaining three are the pairs of irrelevant instruments given by (z_3, z_4), (z_5, z_6), (z_7, z_8). This way there are only a total of 15 options for z_t. While motivated by computational convenience, this scenario is not completely without practical interest. In Euler equation models, it is common for the set of candidate instruments to take the form $\{v_{t-i};\, i = 1, 2 \dots L\}$ where v_t is itself a vector, and then instrument selection problem reduces to deciding which lags of v_t to include in z_t. The results are presented in Tables 3 and 4. For brevity, we break the possible combinations of instruments into four categories: the relevant pair only, the relevant pair plus at least one irrelevant pair, some combination of only the irrelevant pairs, and finally all four pairs. Once again, CCBIC has the highest probability of selecting just the relevant pair at both sample sizes. CCHQIC also does well, but CCAIC again demonstrates a tendency to include both relevant and irrelevant instruments. In nearly every case, the rejection frequency of the t-statistic is very close to the nominal value with the one exception being when CCAIC is used at $T = 100$.

5 Extensions to non-normality and non-linear models

So far we have concentrated on the normal linear model. In this section we demonstrate that the foregoing results apply for non–normal data and also show how to extend the method to nonlinear models.

Inspection of the proofs of Theorems 3 and 4 indicates that the result rests on the following two properties of the likelihood ratio statistic in (11): (i) if z_{2t} is relevant then $T^{-1}LR_T \xrightarrow{p} k > 0$; (ii) if z_{2t} is irrelevant then $LR_T = O_p(1)$. While normality is sufficient for these results, it is not necessary. So we now replace Assumption 1(i) by the following condition.

Assumption 7 *(i)* $\{v_t\}$ *is an i.i.d. process; (ii)* $T^{-1}\sum_{t=1}^{T} v_t v_t' \xrightarrow{p} \Sigma_v = E[v_t v_t']$.

Certain other conditions are needed and for brevity these are listed in the Appendix. Under these conditions, we have the following extension of Theorems 3 and 4.

Theorem 5 *If Assumptions 1 (ii)-(iv), 4(ii), 5, 7, A.1 (in the appendix) and either 4(i) or 6 then:* $\tilde{S} \xrightarrow{p} S^0$.

We now consider the extension to nonlinear models and so must modify our definition of $u_t(\theta)$.

Assumption 8 *Let* $u_t(\theta) = u(w_t, \theta_0)$ *where: (i)* $(r \times 1)$ *random vectors* $\{w_t; -\infty < t < \infty\}$ *form an i.i.d. sequence with sample space* $\mathbf{W} \subseteq \Re^r$*; (ii)* $u : \mathbf{W} \times \Theta \to \Re$ *is continuous on* Θ *for each* $v \in \mathbf{V}$*; (iii)* $E[Z_t u_t(, \theta)]$ *exists and is finite for every* $\theta \in \Theta$*; (iv)* $E[Z_t u_t(\theta)]$ *is continuous on* Θ.

We assume the derivative of u_t, $d_t(\theta) = \partial u_t(\theta)/\partial\theta$ possesses the following properties.

Assumption 9 *(i)* $d_t(\theta)$ *exists and is continuous on* Θ *for each* $w \in \mathbf{W}$*; (ii)* θ_0 *is an interior point of* Θ*; (iii)* $E[\partial f(v_t, \theta_0)/\partial\theta']$ *exists and is finite.*

We also impose the the following restrictions on the evolution of $v_t' = [d_t(\theta_0)', Z_t', u_t(\theta_0)]$.

Assumption 10 *(i)* $\{v_t\}$ *is an i.i.d. process; (ii)* $T^{-1}\sum_{t=1}^{T} v_t v_t' \xrightarrow{p} \Sigma_v = E[v_t v_t']$*; (iii)*

$$\Sigma = \begin{bmatrix} \Sigma_{dd} & \Sigma_{dz} & \Sigma_{xu} \\ \Sigma_{zd} & \Sigma_{zz} & 0_{q_{max}\times 1} \\ \Sigma_{ud} & 0_{1\times q_{max}} & \sigma^2 \end{bmatrix}$$

(iv) Σ_{dd}, Σ_{zz} *are respectively* $p \times p$, $q_{max} \times q_{max}$ *nonsingular matrices, and* σ^2 *is a positive scalar; (v)* $rank\{\Sigma_{dz}\} = p$.

Let $\hat{\theta}_T$ denote the two step GIV estimator based on $E[z_t u_t(\theta_0)] = 0$ where $z_t = S_q Z_t$. It can be shown that under Assumptions 8–10 (and certain other regularity conditions)[10]

[10] *E.g.* See Newey and McFadden (1994).

that

$$T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(0, \sigma^2[\Sigma_{dz}S_q'(S_q\Sigma_{zz}S_q')^{-1}S_q\Sigma_{zd}]^{-1}\right) \tag{14}$$

A comparison of (4) and (14) reveals that d_t plays the same role in nonlinear models as x_t did in the linear model. This motivates the following extensions of our definitions of irrelevance and relevance.

Definition 6 *Irrelevant instrument in nonlinear models*

$z_{t,j}$ *is said to be irrelevant (for the estimation of* θ_0*) if*

$$\rho_i\left[d_t(\theta_0) : \{z_{t,j}, S_{j,q}Z_t\}\right] = \rho_i\left[d_t(\theta_0) : S_{j,q}Z_t\right]$$

for all $i = 1, 2 \ldots p$*, all* $q = p, p+1, \ldots q_{max} - 1$ *and all* $S_{j,q}$*.*

Definition 7 *Relevant instrument in nonlinear models*

$z_{t,j}$ *is said to be relevant if*

$$\rho_i\left[d_t(\theta_0) : \{z_{t,j}, S_{j,q}Z_t\}\right] \geq \rho_i\left[d_t(\theta_0) : S_{j,q}Z_t\right]$$

for all $i = 1, 2 \ldots p$*, all* $q = p, p+1, \ldots q_{max} - 1$ *and all* $S_{j,q}$*, and*

$$\rho_i\left[d_t(\theta_0)x_t : \{z_{t,j}, S_{j,q}Z_t\}\right] > \rho_i\left[d_t(\theta_0) : S_{j,q}Z_t\right]$$

for some i*, all* $q = p, p+1, \ldots q_{max} - 1$ *and all* $S_{j,q}$*.*

The CCIC will be given by (12), where r_i are the sample canonical correlations between $d_t(\tilde{\theta}_T)$ and z_t where $\tilde{\theta}_T$ is an estimator of θ_0. This estimator can be a first or second step estimator but it must satisfy the following condition.

Assumption 11 $\tilde{\theta}_T - \theta_0 = O_p(T^{-1/2})$.

Once again we partition the instrument vector as in Assumption 4(i) and define S^0 by Assumption 4(ii). Let $\tilde{S}$ and $\hat{S}$ be the selection matrices defined by (13) and Definition 5 respectively with $d_t(\tilde{\theta}_T)$ substituted for x_t. The following theorem extends Theorem 3 and Corollary 2 to nonlinear models.

Theorem 6 *If Assumptions 8–11 and A.2 (given in the Appendix) hold then:* $\tilde{S} \xrightarrow{p} S^0$ *and* $\hat{S} \xrightarrow{p} S^0$.

A similar extension is possible for the case where the irrelevant instruments are replaced by nearly irrelevant instruments. However, we omit the details for brevity.

6 Instrument selection in practice

The foregoing analysis has been premised on the assumption that all members of the candidate set are orthogonal to $u_t(\theta_0)$. In practice, this condition is unlikely to be satisfied and so it is desirable to develop a method which selects instruments on the basis of both their orthogonality and relevance. Neither Andrews's (1999) method nor the CCIC approach can meet this objective on their own because each addresses just one of these properties. However, intuition suggests that a combination of the two methods should achieve the desired goal. This section explores the properties of such a selection strategy.

For simplicty, we return to the normal linear model discussed in Sections 2–4. However this time the candidate set of instruments is assumed to have the following composition.

Assumption 12 *(i)* $Z_t = [Z_t(O)', Z_t(C)']'$ *where* $Z_t(O)$ *is a* $q_O \times 1$ *vector of instruments which are orthogonal to* u_t*, that is* $E[Z_t(O)u_t(\theta_0)] = 0$*, and* $Z_t(C)$ *is* $q_C \times 1$ *vector of instruments which are correlated with* $u_t(\theta_0)$*, that is* $E[Z_t(O)u_t(\theta_0)] \neq 0$*; (ii)* $Z_t(O) = [Z_t(R)', Z_t(I)']'$ *where* $Z_t(R)$ *is a* $q_R \times 1$ *vector of instruments which are relevant for the estimation of* θ_0 *and* $Z_t(I)$ *is* $q_I \times 1$ *vector of instruments which are irrelevant for the estimation of* θ_0*; (iii)* S^1 *is the* $q_O \times q_{max}$ *selection matrix which satisfies* $Z_t(O) = S^1 Z_t$*; (iv)* S^0 *is the* $q_R \times q_{max}$ *selection matrix which satisfies* $Z_t(R) = S^0 Z_t$.

So the candidate set now involves the invalid instruments $Z_t(C)$, relevant instruments $Z_t(R)$ and the irrelevant instruments $Z_t(I)$.

For this model Andrews's (1999) criteria takes the form

$$MSC(S_q) \;=\; J_T(S_q) \;-\; K_T(q,p) \tag{15}$$

where $J_T(S_q)$ is the overidentifying restrictions test associated with estimation based on $E[S_q Z_t u_t(\theta_0)] = 0$ and $K_T(q,p)$ is a bonus term dependent on the number of overidentifying restrictions. More specifically, these two components of (15) are given by:

$$J_T(S_q) \;=\; u(\hat{\theta}_T)' Z S_q^{'} (S_q Z' Z S_q^{'})^{-1} S_q Z' u(\hat{\theta}_T)/\hat{\sigma}_T^2$$

where $u(\theta)$ is the $T \times 1$ vector with t^{th} element $u_t(\theta)$, $\hat{\theta}_T$ is defined in (3) and $\hat{\sigma}_T^2 = T^{-1} u(\hat{\theta}_T)' u(\hat{\theta}_T)$. Andrews (1999) considers various choices of bonus term but for simplicity we consider just one, $K_T(q,p) = (q-p)ln(T)$, which uses the bonus associated with the Schwarz criterion. The researcher chooses S_q to minimize $MSC(S_q)$ over $\mathcal{S}$ and so the resulting instrument vector is $\hat{Z}_t(O) = \bar{S} Z_t$ where

$$\bar{S} \;=\; argmin_{S_q \in \mathcal{S}} MSC(S_q)$$

Andrews (1999) provides a set of conditions under which $\bar{S} \xrightarrow{p} S^1$. Therefore, the method selects only those instruments which are orthogonal with probability one in the limit. Once invalid instruments have been excluded, it is then possible to apply CCIC to select which members of $\hat{Z}_t(O)$ are relevant. Our earlier analysis is easily extended to show that such a selection strategy leads to $S^0 Z_t = Z_t(R)$ with probability one in the limit.

We conclude this section by considering how well such a selection strategy works in practice. The simulation design is the same as in Section 4 except that z_7 and z_8 are now endogenous. We report results for three slection strategies: MSC by itself, CCBIC by itself and MSC followed by CCBIC. The results are given in Table 4. For ease of presentation, we divide the possible instrument combinations into four groups: only the relevant instruments, (z_1, z_2); all the orthogonal instruments, $\{z_i;\, i = 1, 2 \ldots 6\}$; only the relevant and endogenous instruments, $\{z_i;\, i = 1, 2, 7, 8\}$; all other combinations. We now

consider the results for each selection strategy in turn. If MSC is used alone then all the orthogonal instruments are selected with high probability. If CCBIC is applied alone then it selects both the relevant and endogenous instruments with high probability. However, if MSC and CCBIC are applied sequentially then the set of relevant instruments is selected with very high probability. At both sample sizes, it is the sequential strategy which leads to a t-statistic whose behaviour is closest to that predicted by asymptotic theory.[11] At $T = 500$ this approximation is good, but at $T = 100$ there is substantial distortion. This contrasts with the behaviour reported in Table 3 for the case in which all the instruments are orthogonal. This suggests that the sequential method selects endogenous instruments with sufficient frequency at $T = 100$ to distort the behaviour of the t-statistic.

7 Concluding remarks

In this paper we have proposed a method for selecting the relevant instruments from a set of orthogonal instruments based on a canonical correlations information criterion. It is shown that the method is consistent and also performs well in finite samples. In practice, it is typically unknown whether potential instruments are orthogonal, and so we propose a sequential strategy. On the first step, Andrews (1999) information criterion based on the overidentifying restrictions test is used to screen out the invalid instruments, and then on the second step the CCIC is applied to determine which of the orthogonal instruments are relevant. Our simulation evidence suggest that this sequential strategy works well.

The consistency proof for CCIC is premised on the assumption that the candidate set consists of instruments which are either relevant or (nearly) irrelevant. This condition serves as an identification condition because it implies CCIC is uniquely minimized by $z_t = Z_t(R)$. While there are certainly cases in which the candidate set has this structure,

[11] It should be noted that within this design the scope for the inclusion of irrelevant instruments is rather limited, and it is for this reason that the use of MSC alone does not create the size distortions reported in Table 1.

it is also desirable to extend the analysis to allow for the case in which the candidate set contains instruments, z_{2t}, which are redundant for some z_{1t} but not for others. Within this more general framework, an important issue is whether or not there is a unique instrument vector which minimizes CCIC. Uniqueness (or identification) is important because of the potential impact of instrument selection on subsequent inferences. If the method selects a unique instrument vector with probability one then inference can proceed as if the chosen vector had been picked *a priori* without recourse to the data. If identification fails then the statistical properties of the instrument selection method may impact on the limiting distribution of the IV estimator. Our decomposition of the candidate set provides one approach to identification. One alternative option is to assume the uniqueness of S^0 directly as follows.

Assumption 13 *(i)* $Z_t = [(S^0 Z_t)', Z_t(NI)']'$ *where* S^0 *is a* $\tilde{q} \times q_{max}$ *selection matrix; (ii)* $\sum_{i=1}^{p} ln\{1 - \rho_i^2[x_t; S^0 Z_t]\} < \sum_{i=1}^{p} ln\{1 - \rho_i^2[x_t; S_q Z_t]\}$ *for all* S_q *with* $q < \tilde{q}$ *or* $q = \tilde{q}$ *and* $S_q \neq S^0$.

It can be recognized that Assumption 13 implies $S^0 Z_t$ minimizes $\sum_{i=1}^{p} ln\{1 - \rho_i^2[x_t; S_q Z_t]\}$ over $\mathcal{S}$ and all other instrument vectors which achieve this minimum contain more elements. This is sufficient to achieve identification, and it is straightforward to establish Theorems 3–6 under this alternative condition.[12] This identification assumption is similar in spirit to Assumptions IDc^0 and $ID\theta_0$ which underlies the consistency results in Andrews (1999). However, in both contexts, such identification conditions are assumptions and it is important to explore the consequences of their failure. This topic is an interesting area for future research.

There are many other interesting extensions of our results. First it is desirable to extend the CCIC method to independently but non–identically distributed data and also dependent data. Second, and more generally, it is interesting to expand the ideas in this

[12]Similar results apply if $Z_t(NI)$ are replaced by $Z_t(I)$ in Assumption 13.

paper to tailor instrument selection to the objectives of the researcher whatever they may be. Implicit in our approach is the assumption that the primary objective of estimation is to perform inference about the unknown parameter vector. Often this is the objective, but in many other cases the objective is something different such as forecasting. In such circumstances, it seems desirable to develop an instrument selection criterion which reflects the ultimate objective of the estimation. All these topics are currently under investigation.

Mathematical Appendix

Proof of Theorem 1

Let $\{\rho_i, i = 1, 2, \ldots, p\}$ be the canonical correlations between x and z and let $\{r_i, i = 1, 2, \ldots, p\}$ be the canonical correlations between x and z_1. To establish the result stated in the theorem, it is necessary to show that the condition for redundancy in (7) implies

$$\{r_i, i = 1, 2, \ldots, p\} = \{\rho_i, i = 1, 2, \ldots, p\}$$

The canonical correlations between x and z_1 are the square roots of the eigenvalues of the matrix

$$\Sigma_{xx}^{-1}\Sigma_{x1}\Sigma_{11}^{-1}\Sigma_{1x} \tag{16}$$

The canonical correlations between x and z are the square roots of the eigenvalues of the matrix

$$\Sigma_{xx}^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx} = \Sigma_{xx}^{-1}\begin{bmatrix} \Sigma_{x1} & \Sigma_{x2} \end{bmatrix}\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1}\begin{bmatrix} \Sigma_{1x} \\ \Sigma_{2x} \end{bmatrix} \tag{17}$$

Using the partitioned inversion result from Magnus and Neudecker (1991)[p. 11] it follows that

$$\Sigma_{zz}^{-1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{11}^{-1}\left(I + \Sigma_{12}F_2\Sigma_{21}\Sigma_{11}^{-1}\right) & -\Sigma_{11}^{-1}\Sigma_{12}F_2 \\ -F_2\Sigma_{21}\Sigma_{11}^{-1} & F_2 \end{bmatrix} \tag{18}$$

where $F_2 = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$. The substitution of (18) into (17) yields

$$\begin{aligned} \Sigma_{xx}^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx} &= \Sigma_{xx}^{-1}\Sigma_{x1}\Sigma_{11}^{-1}\left(I + \Sigma_{12}F_2\Sigma_{21}\Sigma_{11}^{-1}\right)\Sigma_{1x} \\ &\quad -\Sigma_{xx}^{-1}\Sigma_{x1}\Sigma_{11}^{-1}\Sigma_{12}F_2\Sigma_{2x} - \Sigma_{xx}^{-1}\Sigma_{x2}F_2\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{1x} \\ &\quad +\Sigma_{xx}^{-1}\Sigma_{x2}F_2\Sigma_{2x} \end{aligned} \tag{19}$$

The subsitution of (7) in (19) yields

$$\begin{aligned} \Sigma_{xx}^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx} &= \Sigma_{xx}^{-1}\Sigma_{x1}\Sigma_{11}^{-1}\Sigma_{1x} + \Sigma_{xx}^{-1}\Sigma_{x1}\Sigma_{11}^{-1}\Sigma_{12}F_2\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{1x} \\ &\quad -\Sigma_{xx}^{-1}\Sigma_{x1}\Sigma_{11}^{-1}\Sigma_{12}F_2\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{1x} - \Sigma_{xx}^{-1}\Sigma_{x2}F_2\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{1x} \end{aligned}$$

$$
\begin{aligned}
& + \Sigma_{xx}^{-1}\Sigma_{x2}F_2\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{1x} \\
= \;& \Sigma_{xx}^{-1}\Sigma_{x1}\Sigma_{11}^{-1}\Sigma_{1x}
\end{aligned}
$$

which establishes the desired result.

Proof of Theorem 2

Part (i) is proved by Anderson (1984)[p.317]. To establish part (ii) notice that:

$$
T^{-1}LR_T \xrightarrow{p} \sum_{i=1}^{p} \ln\left\{ \frac{1-\rho_i^2[x_t : z_{1t}]}{1-\rho_i^2[x_t : (z_{1t}, z_{2t})]} \right\} = k, \text{ say.} \tag{20}
$$

Now if z_{2t} is not redundant given z_{1t} then it follows that

$$
\begin{aligned}
\rho_i[x_t : (z_{1t}, z_{2t})] &\geq \rho_i[x_t : (z_{1t})] \qquad i = 1, 2, \ldots p \\
\rho_i[x_t : (z_{1t}, z_{2t})] &> \rho_i[x_t : (z_{1t})] \qquad \text{for some } i
\end{aligned}
$$

and so $k > 0$.

Proof of Theorem 3:

The proof rests on considering the limiting behaviour of $CCIC(S_q)$ when the instrument vector is expanded from $z_t = z_{1t}$ to $z_t = [z_{1t}', z_{2t}']'$ and z_t contains $Z_t(R)$. Notice the latter condition implies that (8) is correctly specified. Let $\{r_i;\ i = 1, 2 \ldots p\}$ be the sample canonical correlations between x_t and $z_{1t} = S_{q_1}Z_t$, $\{\hat{\rho}_i;\ i = 1, 2 \ldots p\}$ be the sample canonical correlations between x_t and $z_t = S_q Z_t$ and $q_2 = q - q_1$ then

$$
\begin{aligned}
\Delta &= CCIC(S_q) \;-\; CCIC(S_{q_1}) \\
&= \left[\sum_{i=1}^{p} \ln\left(1-\hat{\rho}_i^2\right) - \sum_{i=1}^{p} \ln\left(1-r_i^2\right) \right] \;+\; q_2 \frac{f(T)}{T} \\
&= -T^{-1}LR_T \;+\; q_2 \frac{f(T)}{T}
\end{aligned} \tag{21}
$$

where LR_T is the likelihood ratio statistic for the null hypothesis that z_{2t} is redundant given z_{1t}.

First consider the case in which z_{2t} is relevant. In this case, it follows from Definitions 2 and 4 that z_{2t} is not redundant given z_{1t} and so from Theorem 2(ii)

$$
\sum_{i=1}^{p} \ln \frac{1-\rho_i^2\,(x_t, z_t)}{1-\rho_i^2\,(x_t, z_{1t})} \;=\; -k \;<\; 0 \tag{22}
$$

Since $f(T)/T \to 0$ as $T \to \infty$, it follows that $\Delta \xrightarrow{p} -k < 0$. Equation (22) has the following important implication. To explore this implication, it is necessary to define $Z_t(R) = [Z_{1t}(R)', Z_{2t}(R)']'$, $Z_t(I) = [Z_{1t}(I)', Z_{2t}(I)']'$ and S^a, S^b be the selection matrices such that $S^a Z_t = [Z_{1t}(R)', Z_{1t}(I)']'$, $S^b Z_t = [Z_t(R)', Z_{1t}(I)']'$. Now (22) implies that

$$\lim_{T\to\infty} P[CCIC(S^b) < CCIC(S^a)] \;=\; 1 \tag{23}$$

and so if the instrument vector contains only some of the relevant instruments then the criterion is always reduced (asymptotically) by augmenting the instrument vector with the omitted relevant instruments.

Now consider the case in which z_{2t} is irrelevant. In this case, $\Delta = o_p(1)$ and so we consider

$$T\Delta \;=\; -LR_T \;+\; q_2 f(T)$$

If z_{2t} is irrelevant then it follows from Definitions 1 and 4 that z_{2t} is redundant given z_{1t}, and so $-LR_T = O_p(1)$ from Theorem 2(i). From Assumption 4, it follows that

$$\lim_{T\to\infty} P[\,\frac{T\Delta}{f\,(T)} = q_2] \;=\; 1 \tag{24}$$

and so $T\Delta$ diverges at rate $f(T)$. Therefore, CCIC increases with probability one as $T \to \infty$ when irrelevant instruments are added to the instrument vector. Equation (24) implies that

$$\lim_{T\to\infty} P[CCIC(S^0) < CCIC(S^b)] \;=\; 1 \tag{25}$$

where S^b is defined above. Equations (23)–(25) imply the desired result.

Proof of Theorem 4

We follow the same basic strategy as the proof of Theorem 3. First notice that if z_{2t} is relevant then we can use the same logic as in the previous proof to deduce (22). So we now focus on proving that (24) holds if z_{2t} is nearly irrelevant. To this end it is useful to introduce the following definition.

Definition A.1 *Near redundancy*

z_{2t} is said to be nearly redundant given z_{1t} if

$$\rho[x_t : z_t] \;=\; \rho[x_t : z_{1t}] \;+\; T^{-1/2}\eta_i \tag{26}$$

for $i = 1, 2 \ldots p$ and $\eta_i > 0$ for some i where $z_t = [z_{1t}^{'}, z_{2t}^{'}]^{'}$.

Now if z_{2t} is nearly irrelevant then it follows from Definition 3 and (26) that is also nearly redundant given any z_{1t}. The statement that z_{2t} is nearly redundant given z_{1t} is equivalent to the restriction that $\Pi_2 = T^{-1/2}C$ in (9) where C is a $p \times q_2$ matrix of finite constants. Recall that LR_T is the likelihood ratio test for $H_0 : \Pi_2 = 0$ and so from standard likelihood theory it follows that LR_T converges to a noncentral χ^2 distribution under the sequence of local alternatives $\Pi_2 = T^{-1/2}C$. Therefore, $LR_T = O_p(1)$ if z_{2t} is nearly irrelevant and so (24) holds.

Proof of Theorem 5

In order to present Assumption A.1 it is necessary to introduce the following definitions. Let:

- $\phi = (vech\{\text{- }\}', vec\{\Pi\}')'$ be a $m \times 1$ vector of unknown parameters in the model (8) and ϕ_0 denote the true value of ϕ.
- $l_t(\phi)$ be the conditional log likelihood for the t^{th} observation.
- $L_T(\phi) = \sum_{t=1}^{T} l_t(\phi)$.
- $vec(\Pi_2) = r(\phi)$.
- $\hat{\phi}_T$ is the unrestricted ML estimator.
- $\tilde{\phi}_T$ is the restricted ML estimator obtained by maximizing $L_T(\phi)$ subject to $r(\phi) = 0$.
- $A_0 = -E\left[\frac{\partial^2 l_t(\phi_0)}{\partial\phi\partial\phi'}\right]$.

- $H_T(\phi_1, \phi_2, \omega)$ be the $m \times m$ matrix with i^{th} row equal to the i^{th} row of $\frac{\partial^2 L_T\left(\bar{\phi}^{(i)}\right)}{\partial \phi \partial \phi'}$, where $\bar{\phi}^{(i)} = \omega^{(i)}\phi_1 + (1-\omega^{(i)})\phi_2$, and $\omega = [\omega^{(1)}, \ldots \omega^{(m)}]$ and $\omega^{(i)}$ lies on the closed unit interval for all i.

Assumption A.1 *(i)* $r(\phi_0) = 0$*; (ii)* $\tilde{\phi}_T \xrightarrow{p} \phi_0$, $\hat{\phi}_T \xrightarrow{p} \phi_0$*; (iii)* $T^{-1}H_T(\phi_1, \phi_2, \omega_T) \xrightarrow{p} -A_0$ *for* $(\phi_1, \phi_2) = (\hat{\phi}_T, \tilde{\phi}_T), (\hat{\phi}_T, \phi_0), (\tilde{\phi}_T, \phi_0)$;[13] *(iv)* $T^{-1/2}\partial L_T(\phi_0)/\partial\phi \xrightarrow{d} N(0, V)$.

Notice that we can use essentially the same structure of proof as Theorem 3 once it is established that (i) if z_{2t} is irrelevant then $LR_T = O_p(1)$; (ii) if z_{2t} is relevant then $LR_T \xrightarrow{p} k > 0$. We first show (i) and then (ii).

Proof of (i):

A second order Taylor expansion for $L_T\left(\tilde{\phi}_T\right)$ around $\hat{\phi}_T$ yields:

$$L_T\left(\tilde{\phi}_T\right) = L_T\left(\hat{\phi}_T\right) + \frac{\partial L_T\left(\hat{\phi}_T\right)}{\partial \phi'}\left(\tilde{\phi}_T - \hat{\phi}_T\right) + 0.5\left(\tilde{\phi}_T - \hat{\phi}_T\right)' H_T(\hat{\phi}_T, \tilde{\phi}_T, \omega_T)\left(\tilde{\phi}_T - \hat{\phi}_T\right) \tag{27}$$

The first order conditions for unrestricted estimation imply that $\frac{\partial L_T\left(\hat{\phi}_T\right)}{\partial \phi'} = 0$ and so (27) becomes

$$L_T\left(\tilde{\phi}_T\right) = L_T\left(\hat{\phi}_T\right) + 0.5\left(\tilde{\phi}_T - \hat{\phi}_T\right)' H_T(\hat{\phi}_T, \tilde{\phi}_T, \omega_T)\left(\tilde{\phi}_T - \hat{\phi}_T\right)$$

With some rearrangement we obtain

$$LR_T = -T^{1/2}\left(\tilde{\phi}_T - \hat{\phi}_T\right)' T^{-1} H_T(\hat{\phi}_T, \tilde{\phi}_T, \omega_T) T^{1/2}\left(\tilde{\phi}_T - \hat{\phi}_T\right) \tag{28}$$

From Assumption A.1 (iii)

$$T^{-1}H_T(\hat{\phi}_T, \tilde{\phi}_T, \omega_T) \xrightarrow{p} -A_0 \tag{29}$$

Now consider $T^{1/2}\left(\tilde{\phi}_T - \hat{\phi}_T\right)$. Since

$$T^{1/2}\left(\tilde{\phi}_T - \hat{\phi}_T\right) = T^{1/2}\left(\tilde{\phi}_T - \phi_0\right) - T^{1/2}\left(\hat{\phi}_T - \phi_0\right) \tag{30}$$

[13] For notational brevity we have used the same ω_T for each combination of (ϕ_1, ϕ_2) but in general the three matrices are evaluted at different ω_T.

we consider each term on the right hand side of (30) in turn.

Using a first order Taylor expansion on the first order conditions we obtain

$$0 \;=\; \frac{\partial L_T\left(\hat{\phi}_T\right)}{\partial \phi} \;=\; \frac{\partial L_T\left(\phi_0\right)}{\partial \phi} + H_T(\hat{\phi}_T, \phi_0, \omega_T)\left(\hat{\phi}_T - \phi_0\right) \tag{31}$$

With some rearrangement (31) becomes

$$T^{1/2}\left(\hat{\phi}_T - \phi_0\right) \;=\; -(T^{-1}H_T(\hat{\phi}_T, \phi_0, \omega_T))^{-1}T^{-1/2}\frac{\partial L_T\left(\phi_0\right)}{\partial \phi} \tag{32}$$

It follows from (32), Assumption A.1(iii)-(iv) that

$$T^{1/2}\left(\hat{\phi}_T - \phi_0\right) \;=\; O_p(1) \tag{33}$$

Now consider $T^{1/2}\left(\tilde{\phi}_T - \phi_0\right)$. The Lagrangean is:

$$\mathcal{L} = L_T\left(\phi\right) - \rho' r\left(\phi\right)$$

and the associated first order conditions are:

$$0 \;=\; \frac{\partial L_T\left(\tilde{\phi}_T\right)}{\partial \phi} - R'\tilde{\rho}_T \tag{34}$$

$$0 \;=\; r\left(\tilde{\phi}_T\right) \tag{35}$$

where $R = \partial r\left(\phi\right)/\partial \phi'$ is independent of ϕ because the restrictions are linear. To proceed we must take two more expansions

$$\frac{\partial L_T(\tilde{\phi}_T)}{\partial \phi} \;=\; \frac{\partial L_T(\phi_0)}{\partial \phi} \;+\; H_T(\tilde{\phi}_T, \phi_0, \omega_T)(\tilde{\phi}_T - \phi_0) \tag{36}$$

$$r(\tilde{\phi}_T) \;=\; r\left(\phi_0\right) \;+\; R(\tilde{\phi}_T - \phi_0) \tag{37}$$

Using $r(\phi_0) = 0$ and (36)-(37) in (34)-(35), it follows that

$$0 = \begin{bmatrix} H_T(\tilde{\phi}_T, \phi_0, \omega_T) & -R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \tilde{\phi}_T - \phi_0 \\ \tilde{\rho}_T \end{bmatrix} + \begin{bmatrix} \partial L_T(\phi_0)/\partial \phi \\ 0 \end{bmatrix} \tag{38}$$

If both sides of (38) are scaled by $T^{1/2}$ and then it follows from (38) that[14]

$$T^{1/2}\left(\tilde{\phi}_T - \phi_0\right) = \left\{B_T^{-1} - B_T^{-1}R'\left[RB_T^{-1}R'\right]^{-1}RB_T^{-1}\right\}T^{-\frac{1}{2}}\frac{\partial L_T\left(\phi_0\right)}{\partial \phi} \tag{39}$$

[14] *E.g.* See Magnus and Neudecker (1991)[p. 11].

where $B_T = -T^{-1}H_T(\tilde{\phi}_T, \phi_0, \omega_T)$. It follows from Assumptions A.1(iii)-(iv) and (39) that

$$T^{1/2}(\tilde{\phi}_T - \phi_0) = O_p(1) \tag{40}$$

The desired result then follows (28), (29), (33) and (40).

Proof of (ii):

This follows directly from Assumption 7 because the canonical correlations are continuous functions of the elements of $T^{-1}\sum_{t=1}^{T} v_t v_t^{'}$.

Proof of Theorem 6

Before we present the proof and Assumption A.2, it is necessary to introduce the following definitions. We first generalize the model in (8) as follows. For a given value of θ we define

$$d_t(\theta) = \Pi(\theta)z_t + v_t(\theta) \tag{41}$$

$$= \Pi_1(\theta)z_{1t} + \Pi_2(\theta)z_{2t} + v_t(\theta) \tag{42}$$

where $\Pi(\theta) = [\Pi_1(\theta), \Pi_2(\theta)]$, $\Pi_i(\theta)$ is a $p \times q_i$ matrix, $q_1 + q_2 = q$ and $z_t^{'} = [z_{1t}^{'}, z_{2t}^{'}]$ is partitioned conformabley. Further set - $(\theta) = Var[v_t(\theta)]$. Note that within this framework the condition for z_{2t} to be redundant given z_{1t} is $\Pi_2(\theta_0) = 0$. With this in mind, we generalize our earlier definitions as follows. Let:

- $\phi(\theta) = [vech\{\text{- }(\theta)\}, vec\{\Pi(\theta)\}]$ and $m = dim(\phi(\theta))$;
- $L_T(\phi; \theta)$ be the log-likelihood function in (10) with $x_t = d_t(\theta)$ for a given value of θ.
- $r\left(\phi(\theta)\right) = vec\{\Pi_2(\theta)\} = 0$;
- $\hat{\phi}_T(\theta)$ be the unrestricted ML estimator of ϕ for a given θ;
- $\tilde{\phi}_T(\theta)$ be the restricted ML estimator;
- $LR_T^* = 2\left[L_T\left(\hat{\phi}_T(\tilde{\theta}_T); \tilde{\theta}_T\right) - L_T\left(\tilde{\phi}_T(\tilde{\theta}_T); \tilde{\theta}_T\right)\right]$ be the LR statistic for $H_0 : r\left(\phi(\tilde{\theta}_T)\right) = 0$;

- H_T^* be the $m \times m$ matrix whose i^{th} row is equal to the i^{th} row of $\partial^2 L_T[\bar{\phi}_T^{(i)}(\tilde{\theta}_T); \tilde{\theta}_T]/\partial\phi\partial\phi'$, where $\bar{\phi}_T^{(i)}(\tilde{\theta}_T) = \omega_T^{(i)}\hat{\phi}_T(\tilde{\theta}_T) + (1-\omega_T^{(i)})\tilde{\phi}_T(\tilde{\theta}_T)$, and $\omega_T^{(i)}$ is a constant which takes a value in the closed unit interval for $i = 1, 2, \ldots m$.

- F_T be the $m \times p$ matrix whose i^{th} row is equal to the i^{th} row of $\partial^2 L_T[\phi(\theta_0); \bar{\theta}_T^{(i)}]/\partial\phi\partial\theta'$, where $\bar{\theta}_T^{(i)} = \omega_T^{(i)}\tilde{\theta}_T + (1-\omega_T^{(i)})\theta_0$, and $\omega_T^{(i)}$ is a constant which takes a value in the closed unit interval for $i = 1, 2, \ldots p$.

Assumption A.2: *(i)* $r(\phi(\theta_0)) = 0$*; (ii)* $\tilde{\phi}_T(\tilde{\theta}_T) \xrightarrow{p} \phi_0(\theta_0)$ *and* $\hat{\phi}_T(\tilde{\theta}_T) \xrightarrow{p} \phi_0(\theta_0)$*; (iii)* $T^{-1}H_T^* \xrightarrow{p} H$*, a matrix of constants; (iv)* $T^{-1}F_T \xrightarrow{p} F$*, a matrix of constants;*
(v) $T^{-1/2}\partial L_T[\phi_0(\theta_0)); \theta_0]/\partial\phi = O_p(1)$.

The proof of Theorem 6 rests on the following Lemma.

Lemma A.1*If (i) Assumptions 8–11 and A.2 hold; (ii)* $d_t(\theta_0)$ *satisfies (41)-(42) evaluated at* $\theta = \theta_0$ *with* $E[v_t(\theta_0)|z_t] = 0$*; then* $LR_T^* = O_p(1)$.

Proof: A second order Taylor expansion for $L_T\left(\tilde{\phi}_T(\tilde{\theta}_T); \tilde{\theta}_T\right)$ around $\hat{\phi}_T(\tilde{\theta}_T)$ (recognising that the score vector is zero) yields:

$$\begin{aligned} L_T\left(\tilde{\phi}_T\left(\tilde{\theta}_T\right); \tilde{\theta}_T\right) &= L_T\left[\hat{\phi}_T\left(\tilde{\theta}_T\right); \tilde{\theta}_T\right] + \\ & \quad 0.5\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right) - \hat{\phi}_T\left(\tilde{\theta}_T\right)\right]' H_T^*\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right) - \hat{\phi}_T\left(\tilde{\theta}_T\right)\right] \end{aligned}$$

After some rearrangement, it follows that

$$LR_T^* = T^{1/2}\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right) - \hat{\phi}_T\left(\tilde{\theta}_T\right)\right]'\left(-T^{-1}H_T^*\right)T^{1/2}\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right) - \hat{\phi}_T\left(\tilde{\theta}_T\right)\right] \qquad (43)$$

From Assumption A.2(ii) it follows that $T^{-1}H_T^* \xrightarrow{p} H$. Now consider $T^{1/2}\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right) - \hat{\phi}_T\left(\tilde{\theta}_T\right)\right]$. Since

$$T^{1/2}\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right) - \hat{\phi}_T\left(\tilde{\theta}_T\right)\right] = T^{1/2}\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right) - \phi_0\left(\theta_0\right)\right] - T^{1/2}\left[\hat{\phi}_T\left(\tilde{\theta}_T\right) - \phi_0\left(\theta_0\right)\right] \qquad (44)$$

we consider each term on the right hand side of (44) in turn.

Using a first order Taylor expansion for the score we obtain

$$T^{1/2}\left[\hat{\phi}_T\left(\tilde{\theta}_T\right) - \phi_0\left(\theta_0\right)\right] = -\left[T^{-1}\frac{\partial^2 L_T\left[\bar{\phi}_T\left(\tilde{\theta}_T\right); \tilde{\theta}_T\right]}{\partial\phi\partial\phi'}\right]^{-1} T^{-1/2}\frac{\partial L_T\left(\phi_0\left(\theta_0\right); \tilde{\theta}_T\right)}{\partial\phi} \qquad (45)$$

Using Assumption A.2 (ii) it follows that

$$-\left[T^{-1}\frac{\partial^2 L_T\left[\bar{\phi}_T\left(\tilde{\theta}_T\right);\tilde{\theta}_T\right]}{\partial\phi\partial\phi'}\right]^{-1} \xrightarrow{p} H^{-1} \;=\; O(1)$$

Expanding now the second term of the right-hand side of (45) around θ_0 we obtain

$$T^{-1/2}\frac{\partial L_T\left(\phi_0(\theta_0);\tilde{\theta}_T\right)}{\partial\phi} = T^{-1/2}\frac{\partial L_T\left(\phi_0(\theta_0);\theta_0\right)}{\partial\phi} + T^{-1}F_T T^{1/2}\left(\tilde{\theta}_T-\theta_0\right) \tag{46}$$

Now using Assumptions 10 and A.2(iii)-(iv), it can be deduced from (45)–(46) that

$$T^{1/2}\left[\hat{\phi}_T\left(\tilde{\theta}_T\right)-\phi_0\left(\theta_0\right)\right] = O_p(1) \tag{47}$$

For $T^{1/2}\left[\tilde{\phi}_T\left(\tilde{\theta}_T\right)-\phi_0\left(\theta_0\right)\right]$, we need to set up the Lagrangean:

$$\mathcal{L}^* = L_T\left(\phi\left(\tilde{\theta}_T\right);\tilde{\theta}_T\right) \;-\; \rho' r\left(\phi\left(\tilde{\theta}_T\right)\right)$$

The first order conditions are:

$$0 \;=\; \frac{\partial L_T\left(\tilde{\phi}_T\left(\tilde{\theta}_T\right);\tilde{\theta}_T\right)}{\partial\phi} - R'\tilde{\rho}_T \tag{48}$$

$$0 \;=\; r\left(\tilde{\phi}_T\left(\tilde{\theta}_T\right)\right) \tag{49}$$

where $R = \partial r(\phi)/\partial\phi'$. To proceed we must take three more expansions

$$\frac{\partial L_T\left(\tilde{\phi}_T(\tilde{\theta}_T);\tilde{\theta}_T\right)}{\partial\phi} \;=\; \frac{\partial L_T\left(\phi_0(\theta_0);\tilde{\theta}_T\right)}{\partial\phi} + P_T^*\left[\tilde{\phi}_T(\tilde{\theta}_T)-\phi_0(\theta_0)\right] \tag{50}$$

$$\frac{\partial L_T\left(\phi_0(\theta_0);\hat{\theta}_T\right)}{\partial\phi} \;=\; \frac{\partial L_T\left(\phi_0(\theta_0);\theta_0\right)}{\partial\phi} + F_T\left(\tilde{\theta}_T-\theta_0\right) \tag{51}$$

$$r\left(\tilde{\phi}_T(\tilde{\theta}_T)\right) \;=\; r\left(\phi_0(\theta_0)\right) \;+\; R\left[\tilde{\phi}_T(\tilde{\theta}_T)-\phi_0(\theta_0)\right] \tag{52}$$

where P_T^* is defined in a similar way to H_T^*. Using $r\left(\phi_0(\theta_0)\right)=0$ and (50)-(52) in (48)-(49), it follows that

$$\begin{aligned} 0 \;=\; & \begin{bmatrix} P_T^* & -R' \\ R & 0 \end{bmatrix}\begin{bmatrix} \tilde{\phi}_T(\tilde{\theta}_T)-\phi_0(\theta_0) \\ \tilde{\rho}_T \end{bmatrix} \\ & + \begin{bmatrix} \partial L_T\left(\phi_0(\theta_0);\theta_0\right)/\partial\phi + F_T(\tilde{\theta}_T-\theta_0) \\ 0 \end{bmatrix} \end{aligned} \tag{53}$$

Using the partitioned inverse formulae in Magnus and Neudecker (1991)[p.11] and scaling by $T^{1/2}$, it follows from (53) that

$$\begin{aligned} T^{1/2}\left[\tilde{\phi}_T(\tilde{\theta}_T) - \phi_0(\theta_0)\right] &= \{ -(T^{-1}P_T^*)^{-1} + (T^{-1}P_T^*)^{-1}R' \times \\ &\quad \left[R(T^{-1}P_T^*)^{-1}R'\right]^{-1} R(T^{-1}P_T^*)^{-1} \} \times \\ &\quad \left[T^{-1/2}\frac{\partial L_T\left(\phi_0(\theta_0);\theta_0\right)}{\partial \phi} + T^{-1}F_T T^{1/2}(\tilde{\theta}_T - \theta_0)\right] \end{aligned} \quad (54)$$

Using Assumptions 10 and A.2, it can be deuced from (54) that $T^{1/2}\left[\tilde{\phi}_T(\tilde{\theta}_T) - \phi(\theta_0)\right] = O_p(1)$. This result, together with (47), (44) and Assumption 11, implies that $LR_T^* = O_p(1)$.

Proof of Theorem 6: The proof follows similar lines to the proof of Theorem 3. It can be recalled that this proof rests on showing (22) holds if z_{2t} is relevant and (24) holds if z_{2t} is irrelevant. We consider these in turn. Using Definition 7 and

$$\begin{aligned} r_i\left[d_t\left(\tilde{\theta}_T\right), z_t\right] - r_i\left[d_t\left(\theta_0\right), z_t\right] &\xrightarrow{p} 0 \\ \rho_i\left[d_t\left(\tilde{\theta}_T\right), z_t\right] - \rho_i\left[d_t\left(\theta_0\right), z_t\right] &\xrightarrow{p} 0 \end{aligned}$$

it follows that (22) holds if z_{2t} is relevant in the nonlinear case as well. Using Definition 6 and Lemma A.1 it can be shown that (24) holds if z_{2t} is irrelevant in the nonlinear case as well.

References

Akaike, H. (1974). 'A new look at statistical model identification', *IEEE Transactions on Automatic Control*, AC-19(6): 716–723.

Andersen, T. G., and Sorensen, B. E. (1996). 'GMM estimation of a stcohastic volatility model: a monte carlo study', *Journal of Business and Economic Statistics*, 14: 328–352.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. Wiley, New York, NY, U. S. A., 2nd edn.

Andrews, D. W. K. (1999). 'Consistent moment selection procedures for Generalized Method of Moments Estimation', *Econometrica*, 67: 543–564.

Bowden, R. J., and Turkington, D. A. (1984). *Instrumental variables*. Cambridge University Press, Cambridge, U. K.

Breusch, T., Qian, H., Schmidt, P., and Wyhowski, D. (1999). 'Redundancy of moment conditions', *Journal of Econometrics*, 91: 89–111.

Hall, A. R., and Peixe, F. P. M. (1999). 'The mean squared error of the instrumental variables estimator when the disturbance has an elliptical distribution', Discussion paper, Department of Economics, North Carolina State University, Raleigh, NC.

Hall, A. R., Rudebusch, G., and Wilcox, D. (1996). 'Judging instrument relevance in instrumental variables estimation', *International Economic Review*, 37: 283–298.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press, Princeton, NJ, U. S. A.

Hannan, E. J., and Quinn, B. G. (1979). 'The determination of order of an autoregression', *Journal of the Royal Statistical Society, Series B*, 41(2): 190–195.

Hansen, L. P. (1982). 'Large sample properties of Generalized Method of Moments estimators', *Econometrica*, 50: 1029–1054.

Hansen, L. P., and Singleton, K. S. (1982). 'Generalized instrumental variables estimation of nonlinear rational expectations models', *Econometrica*, 50: 1269–1286.

Magnus, J. R., and Neudecker, H. (1991). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York, NY.

Newey, W. K., and McFadden, D. L. (1994). 'Large sample estimation and hypothesis testing', in R. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2113–2247. Elsevier Science Publishers, Amsterdam, The Netherlands.

Schwarz, G. (1978). 'Estimating the dimension of a model', *Annals of Statistics*, 6: 461–464.

Staiger, D., and Stock, J. (1997). 'Instrumental variables regression with weak instruments', *Econometrica*, 65: 557–586.

Stock, J., and Wright, J. (1997). 'Asymptotics for GMM estimators with weak instruments', Discussion paper, Kennedy School of Government, Harvard University, Cambridge, MA.

White, H. (1984). *Asymptotic theory for econometricians*. Academic Press, New York, NY, U. S. A.

Table 1: Consequences of the inclusion of irrelevant instruments

q	$bias$	$rmse$	$tstat$	$size$
1	−0.015	0.232	0.057	0.082
2	0.011	0.144	0.165	0.092
3	0.022	0.142	0.212	0.099
4	0.030	0.142	0.316	0.114
5	0.039	0.144	0.391	0.128
6	0.048	0.145	0.444	0.147
7	0.056	0.148	0.493	0.152
8	0.065	0.147	0.573	0.169
9	0.071	0.148	0.625	0.181
10	0.079	0.150	0.708	0.197

Notes: $bias$ and $rmse$ are the simulated bias and rmse of $\hat{\theta}_T$. $tstat$ denotes the simulated mean of t-statistic for $H_0 : \theta_0 = 0$. $size$ denotes the empirical size of the t-test with nominal size 0.1.

Table 2: Properties of simplified selection strategy

	T=100			T=500		
inst.	*CCAIC*	*CCBIC*	*CCHQIC*	*CCAIC*	*CCBIC*	*CCHQIC*
z_1	1.000	0.995	0.999	1.000	1.000	1.000
z_2	0.999	0.994	0.998	1.000	1.000	1.000
z_3	0.166	0.042	0.091	0.163	0.014	0.063
z_4	0.179	0.044	0.094	0.160	0.013	0.055
z_5	0.171	0.038	0.085	0.161	0.013	0.055
z_6	0.181	0.041	0.095	0.168	0.012	0.063
z_7	0.173	0.040	0.090	0.167	0.012	0.057
z_8	0.177	0.043	0.092	0.160	0.010	0.052
tstat	0.374	0.236	0.300	0.178	0.087	0.126
size10	0.128	0.108	0.116	0.105	0.102	0.102

Notes: The entries for *CCAIC*, *CCBIC* and *CCHQIC* denote the frequency with which a particular instrument is selected using that version of CCIC. *tstat* denotes the mean of the t-statistic for $\theta_0 = 0$. *size*10 denotes the empirical size of the t-statistic when the nominal size is 0.1.

Table 3: Probability of selecting various combinations of instruments in full pairwise strategy

	T=100			T=500		
inst.	*CCAIC*	*CCBIC*	*CCHQIC*	*CCAIC*	*CCBIC*	*CCHQIC*
R	0.620	0.961	0.853	0.642	0.995	0.928
R/I	0.376	0.039	0.146	0.354	0.005	0.072
I	0.000	0.000	0.000	0.000	0.000	0.000
all	0.004	0.000	0.000	0.005	0.000	0.000
tstat	0.311	0.166	0.222	0.146	0.071	0.090
size10	0.122	0.099	0.108	0.104	0.100	0.101

Notes: R denotes the pair of relevant instruments, R/I denotes the cases where at least one of the irrelevant pairs is included, I denotes the cases in which only irrelevant pairs are included, *all* denotes the case in which all instruments are included. The entries for $CCAIC$, $CCBIC$ and $CCHQIC$ denote the frequency with which a particular combination is selected using that version of CCIC. *tstat* denotes the mean of the t-statistic for $\theta_0 = 0$. $size10$ denotes the empirical size of the t-statistic when the nominal size is 0.1.

Table 4: Probability of selecting various combinations of instruments with MSC, CCBIC and sequential strategy

	T=100			T=500		
inst.	*MSC*	*CCBIC*	*Seq.*	*MSC*	*CCBIC*	*Seq.*
R	0.000	0.009	0.863	0.000	0.000	0.989
O	0.872	0.000	0.000	0.989	0.000	0.000
R/E	0.000	0.965	0.022	0.000	0.996	0.000
Others	0.129	0.026	0.115	0.011	0.004	0.012
tstat	0.704	2.365	0.486	0.202	5.082	0.081
size10	0.206	0.720	0.182	0.111	0.998	0.103

Notes: R denotes the pair of relevant instruments, O denotes the case where all the orthogonal pairs are included, R/E denotes the case in which only the relevant and endogenous pairs are included, *Others* denotes all other combinations. The entries for MSC, $CCBIC$ and *Seq.* (sequential) denote the frequency with which a particular combination is selected using each method. *tstat* denotes the mean of the t-statistic for $\theta_0 = 0$. *size*10 denotes the empirical size of the t-statistic when the nominal size is 0.1.