Information in Generalized Method of Moments Estimation and Entropy Based Moment Selection[1]

Alastair R. Hall
North Carolina State University[2]
Atsushi Inoue
North Carolina State University
Kalidas Jana
Trinity University
and
Changmock Shin
North Carolina State University

April 29, 2005

[1]Paper prepared for "A conference in honor of Arnold Zellner: recent developments in the theory, method and application of information and entropy econometrics", American University, Washington DC, September 19-21, 2003. An earlier version of this paper was circulated under the title "A canonical correlations interpretation of Generalized Method of Moments estimation with applications to moment selection". We are extremely grateful to Peter Schmidt, Amos Golan, Yuichi Kitamura and four anonymous referees for useful comments on this work.

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

Abstract

In this paper, we make five contributions to the literature on information and entropy in Generalized Method of Moments (GMM) estimation. First, we introduce the concept of the long run canonical correlations (LRCC's) between the true score vector and the moment function $f(v_t, \theta_0)$ and show that they provide a metric for the information contained in the population moment condition $E[f(v_t, \theta_0)] = 0$. Second, we show that the entropy of the limiting distribution of the GMM estimator can be written in terms of these LRCC's. Third, motivated by the above results, we introduce an information criterion based on this entropy that can be used as a basis for moment selection. Fourth, we introduce the concept of nearly redundant moment conditions and use it to explore the connection between redundancy and weak identification. Fifth, we analyze the behaviour of the aforementioned entropy based moment selection method in two scenarios of interest; these scenarios are: (i) nonlinear dynamic models where the parameter vector is identified by all the combinations of moment conditions considered; (ii) linear static models where the parameter vector may be weakly identified for some of the combinations considered. The first of these contributions rests on a generalized information equality that is proved in the paper, and may be of interest in its own right.

1 Introduction

Generalized Method of Moments (GMM) provides a computationally convenient method for obtaining estimators of the parameters of economic models based on a set of population moment conditions. The resulting estimators can be shown to be consistent and asymptotically normal under fairly weak regularity conditions. In most cases of interest, the researcher is actually faced with a candidate set from which to choose the moment conditions to be used in the estimation. Intuition suggests that the choice should reflect the information content of the moment conditions relative to the desired inferences. However, to date, no metric for information has been proposed in the GMM framework.[1] In this paper, we show that the entropy of the limiting distribution of the GMM estimator provides a metric for information in the moment condition regarding the unknown parameter vector. In view of this property, we propose an information criterion for moment selection based on this entropy and analyze its properties. Our analysis rests on a number of other new results presented here regarding GMM estimators that are of interest in their own right.

An outline of the paper is as follows. In Section 2, we introduce the concept of the long run canonical correlations (LRCC's) between the true score vector and a moment function $f(v_t, \theta_0)$ and show that various results pertaining to the GMM estimator based on these LRCC's. As part of this analysis, we prove a generalized information equality that may be of independent interest. In Section 3, we discuss the entropy of the limiting distribution of the GMM estimator, and show that this entropy can be written in terms of these LRCC's. It is shown that the entropy of the limiting distribution of the GMM estimator provides a metric for the information contained in the moment condition, $E[f(v_t, \theta_0)] = 0$. Motivated by the aforementioned results, we introduce in Section 3 an information criterion based on this entropy that can be used as a basis for moment selection and establish conditions for its

[1] We confine our discussion entirely to estimation within the Classical paradigm. See Zellner (2003) for a survey of available results on Bayesian Method of Moments estimation.

consistency in nonlinear dynamic models. The aforementioned consistency result depends crucially on the assumption that the parameter vector is identified by all the combinations of moment conditions considered. This assumption is standard in the GMM literature on moment selection. However, it is clearly possible that moment selection may involve a choice set in which the parameter vector is weakly identified by some of the combinations of moment conditions considered. In the remainder of the paper, we consider exactly this situation in the context of the linear static model estimated via instrumental variables. To this end, in Section 4, we introduce the concept of nearly redundant moment conditions and use it to explore the connection between redundancy and weak identification. Section 5 then establishes conditions for the consistency of our entropy based method for moment selection when the parameter vector may be weakly identified for some of the combinations considered. Section 6 concludes. All proofs are relegated to a mathematical appendix.

2 GMM and a Generalized Information Equality

In this section, we present a generalized information equality and exploit it to show that certain statistical properties of the GMM estimator can be stated in terms of the long run canonical correlations between the moment function, $f(v_t, \theta_0)$, and the score function, $s_t(\theta_0)$.

Throughout our discussion of GMM in its most general form, we consider the case in which the data satisfy the following condition.

Assumption 1 *$\{v_t \in \mathcal{V}, t = 1, 2, \ldots\}$ is a sequence of strictly stationary and ergodic random vectors where $\mathcal{V} \subseteq \Re^s$.*

We consider the GMM estimator of the unknown $p \times 1$ parameter vector θ_0 based on the population moment condition $E[f(v_t, \theta_0)] = 0$ where $f : \mathcal{V} \times \Theta \to \Re^q$ where $q \geq p$. This estimator is defined to be

$$\hat{\theta}_T(f) = Argmin_{\theta \in \Theta}\, g_T(\theta)' W_T g_T(\theta) \tag{1}$$

where $g_T(\theta) = T^{-1}\sum_{t=1}^{T} f(v_t, \theta)$ and W_T is a positive semi–definite weighting matrix which converges in probability to $S(f)^{-1}$ where

$$S(f) \;=\; \lim_{T\to\infty} Var[T^{1/2} g_T(\theta_0)]\,. \tag{2}$$

Under certain regularity conditions it can be shown that the estimator has the following asymptotic distribution.[2]

Assumption 2 *$T^{1/2}[\hat{\theta}_T(f) - \theta_0] \xrightarrow{d} N(0, V_\theta(f))$ where $V_\theta(f) = [G(f)'S(f)^{-1}G(f)]^{-1}$ and $G(f) = E[\partial f(v_t, \theta_0)/\partial\theta']$.*

Note that a necessary condition for the distributional result in Assumption 2 is that $rank\{G(f)\} = p$; this is commonly termed the condition for local identification.

To present the generalized information equality, we need certain additional definitions and certain regularity conditions. Let $p(v_t|V_{t-1}, \theta_0)$ denote the conditional probability density function (pdf) of v_t given the infinite history of the series $V_{t-1} = (v_{t-1}, v_{t-2}, \ldots)$ and $p(V_t|\theta_0)$ denote the joint distribution of V_t. We impose the following set of primitive conditions.

Assumption 3 *(i) $f : \mathcal{V} \times \Theta \to \Re^q$; (ii) θ_0 is an interior point of the parameter space Θ; (iii) there is a neighborhood of θ_0, $\mathcal{N}_{\theta_0}$, such that, for all $\theta \in \mathcal{N}_{\theta_0}$ and t,*
$\int_{\mathcal{V}_{(-\infty,t)}} f(v_t, \theta)p(V_t|\theta)dV_t = 0$; where $\mathcal{V}_{(-\infty,t)}$ is the sample space for V_t; (iv) $f(v_t, \theta)$ and $p(V_t|\theta)$ are continuously differentiable in θ with probability one; (v) $E[\|f(v_t, \theta_0)\|^\zeta] < \infty$, and $E[\|(\partial/\partial\theta)p(V_t|\theta_0)\|^\eta] < \infty$ where $\zeta > 1$, $\eta > 1$ and $1/\zeta + 1/\eta < 1$; (vi) there are functions $g(\cdot)$ and $\{h_t(\cdot)\}$ such that, for all $\theta \in \mathcal{N}_{\theta_0}$ and t, $\|\partial f(v_t, \theta)/\partial\theta\| \le g(v_t)$, $\|\partial p(V_t|\theta)/\partial\theta\| \le h_t(V_t)$, $\int_{\mathcal{V}_{(-\infty,t)}} g(v_t)h_t(V_t)dV_t < \infty$.

(vii) $\{v_t\}$ is strong mixing with mixing coefficients $\{\alpha_j\}$ satisfying $\sum_{j=0}^{\infty} j\alpha_j^{1-1/\zeta-1/\eta} < \infty$.

It is worth noting that only Assumption 3(ii) and the conditions involving $f(\cdot)$ are typically imposed to deduce the asymptotic distribution given in Assumption 2. However, the

[2] See Hansen (1982) or Hall (2005)[Chapter 3].

remaining conditions in Assumption 3 are not particularly restrictive and are likely to hold in most applications of interest.

Now define $s_t(\theta_0)$ to be the conditional score with respect to θ evaluated at $\theta = \theta_0$, that is $s_t(\theta_0) = \{\partial ln[p(v_t|V_{t-1}, \theta_0)]/\partial\theta\}|_{\theta=\theta_0}$. The generalized information equality is given in the following theorem.

Theorem 1 *(a) If Assumptions 1 and 3(i)-(vi) hold, and $\{v_t\}$ is strong mixing with mixing coefficients $\{\alpha_j\}$ satisfying $\sum_{j=0}^{\infty} \alpha_j^{1-1/\zeta-1/\eta} < \infty$, then*

$$G(f) = -\sum_{n=0}^{\infty} E[f(v_t, \theta_0) s_{t-n}(\theta_0)'] ; \tag{3}$$

(b) If Assumptions 1 and 3 hold then it follows that:

$$G(f) = -lim_{T\to\infty} Cov[T^{-1/2} \sum_{t=1}^{T} f(v_t, \theta_0), T^{-1/2} \sum_{t=1}^{T} s_t(\theta_0)] . \tag{4}$$

Theorem 1 shows that the expected Jacobian of the moment condition is equal to the negative of the long run covariance between the moment function and the (unknown) true score function. This theorem generalizes to stationary time series the information equality presented by Godambe (1960) and Tauchen's (1985)[Theorem 5] for iid data.

We now exploit this generalized information equality to show that certain statistical properties of the GMM estimator can be stated in terms of the long run canonical correlations between the moment function, $f(v_t, \theta_0)$, and the score function, $s_t(\theta_0)$. To our knowledge, the concept of long run canonical correlation is new to the literature and so we first define what is meant by this term in our context here.

Definition 1 *The long run canonical correlations (LRCC) between $f(v_t, \theta_0)$ and $s_t(\theta_0)$ are the canonical correlations between $T^{-1/2} \sum_{t=1}^{T} f(v_t, \theta_0)$ and $T^{-1/2} \sum_{t=1}^{T} s_t(\theta_0)$ and are denoted by $\{\rho_i(f);\ i = 1, 2, \ldots p\}$.*

A more detailed definition of LRCC's is relegated to the appendix for brevity; further discussion can be found in Jana (2005).

The following theorem presents alternative representations for both the condition for local identification and the limiting variance $V_\theta(f)$ in terms of the long run canonical correlations (LRCC's) between $f(v_t, \theta_0)$ and the true score vector.

Theorem 2 *If Assumptions 1 and 3 hold then:*

(i) $rank\{G(f)\}$ *equals the number of non-zero long run canonical correlations between* $f(v_t, \theta_0)$ *and* $s_t(\theta_0)$*;*

(ii) the variance of the limiting distribution of the GMM estimator can be decomposed as follows:

$$V_\theta(f) \;=\; A(f)R^{-2}(f)A(f)'$$

where $R(f) = diag(\rho_1(f), \rho_2(f), \ldots, \rho_p(f))$ *and* $A(f)$ *is the* $p \times p$ *matrix with* i^{th} *column* $a_i(f)$ *where* $a_i(f)$ *is the generalized eigenvector satisfying* $[\,C(f)'S(f)^{-1}C(f) \;-\; \rho_i^2(f)\mathcal{I}_\theta\,]a_i(f) = 0$*, where* $\mathcal{I}_\theta = E[s_t(\theta_0)s_t(\theta_0)']$ *and* $C(f) = lim_{T\to\infty} Cov[T^{-1/2}\sum_{t=1}^T f(v_t, \theta_0), T^{-1/2}\sum_{t=1}^T s_t(\theta_0)]$.

Now suppose that Assumption 3 is satisfied for moment functions $f_j(v_t, \theta_0)$*,* $j = 1, 2$ *and let* $\{\rho_i(f_j)\}$ *be the population LRCC's between* $f_j(v_t, \theta_0)$ *and* $s_t(\theta_0)$*. Define* $\hat{\theta}_T(f_j)$ *to be the GMM estimators based on* $E[f_j(v_t, \theta_0)] = 0$*. Assume that the limiting distribution of* $\hat{\theta}_T(f_j)$ *is given by Assumption 2 with* $f = f_j$*.*

(iii) (a) $V_\theta(f_1) = V_\theta(f_2)$ *if and only if* $\rho_i^2(f_1) = \rho_i^2(f_2)$ *for* $i = 1, 2, \ldots p$*; (b)* $V_\theta(f_1) - V_\theta(f_2)$ *is positive semi-definite if and only if* $\rho_i^2(f_2) \geq \rho_i^2(f_1)$ *for* $i = 1, 2, \ldots p$ *and the inequality is strict for at least one value of* i*.*

Comment 1: Theorem 2(i) provides an interesting perspective on the condition for local identification. The estimator can only be identified if all the LRCC's between the score and $f(v_t, \theta_0)$ are non-zero.

Comment 2: Theorem 2(ii) has some interesting implications for asymptotic efficiency. First, notice that if $\{\rho_i^2(f) = 1; i = 1, 2, \ldots p\}$ then Lemma A.1 (in the appendix) and Theorem 2(ii) imply that $V_\theta(f) = \mathcal{I}_\theta^{-1}$ which is the asymptotic version of the Cramer Rao lower bound for estimation of θ_0. Second, Theorem 2(ii) shows that the MLE is asymptotically efficient relative to the GMM estimator because

$$\begin{aligned} V_\theta^{-1}(s_t(\theta_0)) - V_\theta^{-1}(f) &= E[\sum_{n=-\infty}^{\infty} s_t(\theta_0) s_{t-n}(\theta_0)'] - E[\sum_{n=-\infty}^{\infty} s_{t-n}(\theta_0) f(v_t, \theta_0)'] \\ &\quad \times \{E[\sum_{n=-\infty}^{\infty} f(v_t, \theta_0) f(v_{t-n}\, \theta_0)']\}^{-1} E[\sum_{n=-\infty}^{\infty} f(v_t, \theta_0) s_{t-n}(\theta_0)'] \end{aligned}$$

is the population residual covariance matrix of the spectral regression of $s_t(\theta_0)$ on $f(v_t, \theta_0)$ at frequency zero and thus is always positive semidefinite. This result generalizes to stationary time series the analogous result derived by Godambe (1960) for iid data.

Comment 3: Theorem 2(iii) indicates that the LRCC's are sufficient statistics for efficiency comparions between estimators based on differents moment conditions. An illustration is considered in Comment 4 below.

Comment 4: Breusch, Qian, Schmidt, and Wyhowski (1999) use the term redundancy to describe the situation in which the augmentation of the population moment condition has no effect on the asymptotic variance of the estimator. More specifically, suppose that $f(v_t, \theta) = [f_1(v_t, \theta)', f_2(v_t, \theta)']'$ then $E[f_2(v_t, \theta_0)] = 0$ is said to be redundant for θ_0 given $E[f_1(v_t, \theta_0)] = 0$ if $V_\theta(f) = V_\theta(f_1)$. Therefore, if $E[f_2(v_t, \theta_0)] = 0$ is redundant given $E[f_1(v_t, \theta_0)] = 0$ then it provides no information about θ_0 beyond that already in $E[f_1(v_t, \theta_0)] = 0$. The converse of redundancy is termed *non*-redundancy. If $E[f_2(v_t, \theta_0)] = 0$ is non-redundant given $E[f_1(v_t, \theta_0)] = 0$ then $V_\theta(f_1) - V_\theta(f)$ is positive semi-definite and so $E[f_2(v_t, \theta_0)] = 0$ provides additional information.[3] Theorem 2(iii) implies that redundancy can be categorized using LRCC's between the moment function and score. Specifically, it follows from Theorem 2(iii) that $E[f_2(v_t, \theta_0)] = 0$ is redundant for

[3]Note that the asymptotic variance can never increase as a result of augmenting the population moment condition.

the estimation of θ_0 given $E[f_1(v_t, \theta_0)] = 0$ if and only if $\rho_i^2(f) = \rho_i^2(f_1),\ i = 1, 2, \ldots p$; $E[f_2(v_t, \theta_0)] = 0$ is not redundant for the estimation of θ_0 given $E[f_1(v_t, \theta_0)] = 0$ if and only if: $\rho_i^2(f) \geq \rho_i^2(f_1),\ i = 1, 2, \ldots p$ and $\rho_i^2(f) > \rho_i^2(f_1)$ for at least one i.

Comment 5: While our discussion has been in the context of GMM estimators, it should be noted that the results in Theorem 2 extend to Empirical Likelihood estimators (see Owen 1988, 2001; Qin and Lawless, 1994; Kitamura, 1997), Minimum Chi-Square estimators (Neyman, 1949) and Quadratic Inference Function estimators (Qu, Lindsay, and Li, 2000). This follows because these estimators have the same condition for local identification and limiting distribution as the GMM estimator.

3 Information and Entropy in GMM Estimation

To date, the GMM literature has not yielded a generally accepted metric for the information content of population moment conditions. Instead, the literature has focused on the statistical consequences of particular information scenarios that arise in econometric models. Of these scenarios, three of the most important are: (i) the optimal choice of moment condition, that is the choice which yields maximum information about θ_0; (ii) redundant moment conditions, that is moment conditions which provide no incremental information about θ_0; (iii) weak identification, that is the case where there is insufficient information to yield a consistent estimator of θ_0.[4] In this section, we consider the entropy of the limiting distribution of $\hat{\theta}_T(f)$. It is shown that this entropy can be used to characterize the three information scenarios of interest in GMM estimation and hence provides a continuous measure of the information about θ_0 in $E[f(v_t, \theta_0)] = 0$. This result motivates us to propose an information criterion for moment selection based on this entropy that is also introduced in this section. It should be noted that it has long been recognized that entropy can be used

[4]For references to these scenarios see: for (i) Hall (2005)[Chapter 7]; for (ii) see Comment 4 above; for (iii) Hall (2005)[Chapter 8.2].

as a basis for model selection;[5] the unique aspect of our contribution here is the application of this principle to GMM estimation.

Ahmed and Gokhale (1989) derive the entropy for the normal distribution.[6] Applying their result to the limiting distribution of the GMM estimator of θ_0 based on $E[f(v_t, \theta_0)] = 0$ (given in Assumption 2), it follows that the entropy of this distribution is:

$$ent_\theta(f) \;=\; 0.5p[1 + ln(2\pi)] \;-\; 0.5ln[|G(f)^{'}S(f)^{-1}G(f)|] \tag{5}$$

It follows from Theorem 2(ii) and Lemma A.1 (in the Appendix) that

$$ent_\theta(f) \;=\; 0.5p[1 + ln(2\pi)] \;-\; 0.5\sum_{i=1}^{p} ln[\rho_i^2(f)] \;+\; 0.5ln[|\mathcal{I}_\theta^{-1}|] \tag{6}$$

Notice that the entropy only depends on the choice of moment condition via $\{\rho_i^2(f)\}$. Given this structure, it is immediately apparent that Theorem 2 can be used to characterize the three information scenarios described at the beginning of this section in terms of the entropy.

Corollary 1 *Let $\{v_t\}$ satisfy Assumption 1 and define $\mathcal{F} = \{f(\cdot)$ such that Assumption 2 holds$\}$.*

(i) Let $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ and f^0 be the optimal choice of moment condition from $\tilde{\mathcal{F}}$ in the sense that $V_\theta(f) - V_\theta(f^0)$ is positive semi-definite for all $f \in \tilde{\mathcal{F}}$. Then $ent_\theta(f^0) \leq ent_\theta(f)$ for all $f \in \tilde{\mathcal{F}}$.

(ii) Define $f(v_t, \theta) = [f_1(v_t, \theta)', f_2(v_t, \theta)']'$. Assume $f_i \in \mathcal{F}$ for $i = 1, 2$. If $E[f_2(v_t, \theta_0)] = 0$ is redundant for θ_0 given $E[f_1(v_t, \theta_0)] = 0$ then $ent_\theta(f) = ent_\theta(f_1)$. If $E[f_2(v_t, \theta_0)] = 0$ is non-redundant for θ_0 given $E[f_1(v_t, \theta_0)] = 0$ then $ent_\theta(f) < ent_\theta(f_1)$.

(iii) If $rank\{G(f)\} < p$ and so θ_0 is unidentified by $E[f(v_t, \theta_0)] = 0$ then $ent_\theta(f) = \infty$.

Corollary 1 suggests that the entropy of the limiting distribution provides a measure of the information content of moment conditions in GMM estimation. As such, it would

[5] For example, see the review article by Maasoumi (1993) and the references therein.

[6] The entropy is defined to be the expectation of the log of the probability density function of the distribution.

appear to be a natural basis for selection of moments from a candidate set of moment conditions that are known to be valid.[7]

To consider the problem of moment selection, it is necessary to introduce some additional notation. It is assumed that the candidate set of scalar functions which can form the basis for the population moment condition is finite. It is convenient to stack these scalar functions into a single vector $f_{max}(\cdot)$ whose dimension is denoted by q_{max}. Following Andrews (1999), we use a $q_{max} \times 1$ selection vector c to denote which elements of the candidate set are included in a particular moment condition. We therefore now index $f(\cdot)$ by c. If $c_j = 1$ then the j^{th} element of $f_{max}(\cdot)$ is included in $f(\cdot; c)$, and $c_j = 0$ implies this element is excluded. Note that $|c| = c'c$ equals the number of elements in $f(\cdot; c)$. The set of all possible selection vectors is denoted by C, that is

$$C = \left\{ c \in \Re^{q_{max}};\ c_j = 0, 1,\ \text{for } j = 1, 2, \ldots q_{max},\ \text{ and } \ c = (c_1, \ldots c_{q_{max}})',\ |c| \geq p \right\}.$$

For brevity, statistics of interest are now indexed by c and so $\hat{\theta}_T(c)$ denotes the GMM estimator based on $E[f(v_t, \theta_0; c)] = 0$, and $V_\theta(c)$ denotes the variance of its limiting distribution given in Assumption 2.[8]

It is assumed that the researcher wishes to base the estimation on the subset of the available moment conditions which is asymptotically efficient but contains no redundant moment conditions. Asymptotic efficiency is a standard requirement in statistics, and so, for brevity, we do not justify its merits here. The exclusion of redundant moment conditions is a relatively new criterion, and so deserves some justification. Hall and Peixe (2003) report simulation evidence that the inclusion of redundant moment conditions can lead to a serious deterioration in the quality of the limiting distribution (in Assumption 2) as an approximation to finite sample behaviour.[9] It is these findings that motivate the inclusion

[7] See below for discussion of the case in which some elements of candidate set may be invalid.

[8] Note that $\hat{\theta}_T(c)$ and $V_\theta(c)$ denote $\hat{\theta}_T(f)$ and $V_\theta(f)$ evaluated at $f = f(\cdot; c)$.

[9] Further evidence to this effect is presented below. Earlier studies reported similar findings but, since these studies predated the Breusch, Qian, Schmidt, and Wyhowski's (1999) paper, their conclusions are not

of non-redundancy in the objective of moment selection. For ease of exposition, we use the term "relevant" moment conditions to denote the subset of the available moment conditions which are asymptotically efficient but contain no redundant moment conditions. A formal definition of relevance follows.

Definition 2 *c_r is the selection vector associated with the relevant moment conditions if the following three properties hold: (i) $c_r \in C$; (ii) $V_\theta(\iota_{q_{max}}) = V_\theta(c_r)$ where $\iota_{q_{max}}$ is a $q_{max} \times 1$ vector of ones; (iii) $V_\theta(c_{r,1}) - V_\theta(c_r)$ is positive semi–definite for $c_r = c_{r,1} + c_{r,2}$ and $c_{r,1} \in C$.*

A few observations about this definition are in order. Part (ii) states that estimation based on the complete candidate set and the relevant subset yield the same asymptotic variance for the estimator. Since there is no cost asymptotically to the inclusion of redundant moment conditions, part (ii) implies the asymptotic efficiency of estimation based on the relevant moment conditions. Note here that asymptotic efficiency is relative to all possible choices of moment condition from the candidate set. An implication of this property is that if $c_r \neq \iota_{q_{max}}$ then all remaining elements of the candidate set are redundant given the relevant subset, that is $V_\theta(c_r) = V_\theta(c_r + c_i)$ for $c_r' c_i = 0$ and $(c_r + c_i) \in C$.

To motivate the form of the information criterion proposed below, we note two features of the entropy in (5). First, p - the dimension of θ - is constant across all moments considered and so the only part of $ent_\theta(f)$ that changes with the choice of moments is $-ln[|G(f)'S(f)^{-1}G(f)|] = ln[|V_\theta(f)|]$. Second, Corollary 1 implies that we wish to find the choice of moment condition that minimizes $ent_\theta(f)$ and hence $ln[|V_\theta(f)|]$ across all choices of f in the candidate set. These considerations lead us to consider the following information criterion (where we return to indexing moments by c)

$$RMSC(c) \;=\; ln[|\hat{V}_{\theta,T}(c)|] \;+\; \kappa(|c|, T) \tag{7}$$

couched in terms of "redundancy"; see Hall (2005)[Chapter 6] for further discussion and references.

where $\hat{V}_{\theta,T}(c)$ denotes a consistent estimator of $V_\theta(c)$ and $\kappa(|c|,T)$ is a deterministic penalty that is an increasing function of the number of moments, $|c|$. A natural choice for the covariance matrix estimator is

$$\hat{V}_{\theta,T}(c) = [G_T(\hat{\theta}_T(c);c)'\hat{S}_T^{-1}(c)G_T(\hat{\theta}_T(c);c)]^{-1}$$

where $G_T(\theta;c) = T^{-1}\sum_{t=1}^{T}\partial f(v_t,\theta;c)/\partial\theta'$, $\hat{S}_T(c) \xrightarrow{p} S(c)$ and $S(c) = \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} f(v_t,\theta_0;c)]$. The acronym RMSC stands for *relevant moment selection criterion.*

Our proposal is to base estimation on the selection vector that minimizes the criterion over C, that is the selected vector is given by

$$\hat{c}_T \;=\; \text{argmin}_{c\in C} RMSC(c)$$

To analyze the limiting properties of $\hat{c}_T$, we require certain regularity conditions. We first present these conditions and a consistency result for $\hat{c}_T$, and then present a number of comments regarding the construction of $RMSC(c)$ and its relationship to other information criterion in the literature.

To present these regularity conditions, it is necessary to define the set of selection vectors that are asymptotically efficient relative to the candidate set,

$$\mathcal{C} \;=\; \{c;\, V_\theta(\iota_{q_{max}}) = V_\theta(c),\, c\in C\}$$

and also the subset of $\mathcal{C}$ of minimum length,

$$\mathcal{C}_{min} \;=\; \{c;\, c\in\mathcal{C},\, |c| \le |\bar{c}| \text{ for all } \bar{c}\in\mathcal{C}\}\,.$$

Using this notation, we impose the following conditions.

Assumption 4 *(i) c_r satisfies Definition 2 and $\mathcal{C}_{min} = \{c_r\}$; (ii) $E[f(v_t,\theta_0;c)] = 0$ if and only if $\theta = \theta_0$ for all $c\in C$; (iii) $\hat{V}_{\theta,T}(c) = V_\theta(c) + O_p(\tau_T^{-1})$ where $\tau_T\to\infty$ as $T\to\infty$; (iv) for any $\tilde{c},\bar{c}\in C$ such that $|\bar{c}| > |\tilde{c}|$, $\tau_T[\kappa(|\bar{c}|,T) - \kappa(|\tilde{c}|,T)] \to +\infty$ as $T\to\infty$, and $\kappa(|c|,T) = o(1)$ for every $c\in C$.*

Assumptions 4(i) and 4(ii) are the identification conditions for the relevant moment conditions and the parameters, respectively. When the weighting matrix is the inverse of the sum of a fixed number of autocovariances, $\tau_T = T^{1/2}$. When the weighting matrix is the inverse of a heteroscedasticity autocorrelation covariance (HAC) matrix calculated with bandwidth ℓ_T such that $\ell_T \to \infty$ as $T \to \infty$ and $\ell_T = o(T^{1/2})$ then $\tau_T = (T/\ell_T)^{1/2}$. Andrews (1991) provides more primitive conditions for Assumption 4(iii) for this case (e.g., his Assumptions B and C).

The following theorem shows that $\hat{c}_T$ is consistent for c_r.

Theorem 3 *Under Assumption 4, it follows that:* $\hat{c}_T \xrightarrow{p} c_r$.

We now present certain comments regarding the construction of $RMSC(c)$.

Comment 6: It is useful to highlight the differences between $RMSC$ and the Moment Selection Criterion (MSC) proposed by Andrews (1999). MSC is designed to select which moments out of the candidate set represent valid information. For a given c, $MSC(c)$ is defined to be the overidentifying restrictions test plus a bonus term that is equal to $-\kappa(|c|, T)$ (using our notation). The selected moment condition is chosen by minimizing the criterion over $\mathcal{C}$. As pointed out by Hall and Peixe (2003), one weakness of $MSC(c)$ is that it selects moments only on the basis of their validity and takes no account of their information content. In practice, it may be desirable to use $MSC(c)$ and $RMSC(c)$ sequentially but an exploration of their combined use is beyond the scope of this paper.

Comment 7: A number of information criteria have been proposed for the problem of order selection in time series; for example see Akaike (1974), Hannan and Quinn (1979) and Schwarz (1978). However there are important differences between this problem and the one of moment selection considered here. For example, consider the use of Schwarz's (1978) BIC to select the order of an autoregressive processes (AR). BIC involves estimating an AR process for each possible order and calculating the associated error variance. The chosen order is the one that minimizes the log of the error variance plus the deterministic penalty

$pln(T)/T$ where p is the associated AR order. Note two important differences to our setting here: (i) the sample information in BIC is the error variance of the fitted AR model whereas the sample information in $RMSC$ is the variance of estimated parameters; (ii) in BIC, the dimension of the parameter vector changes with the AR order, but in RMSC the dimension of the parameter vector is constant over all choices of moment condition.

Comment 8: Assumption 4(iv) places rather general conditions on the deterministic penalty term and is modelled on Andrews's (1999) Assumption MSC that underpins his analysis of MSC. Examples of penalty terms that satisfy this condition are: $\kappa(|c|,T) = (|c|-p)\ln(\tau_T)/\tau_T$ (BIC-type penalty) and $\kappa(|c|,T) = (|c|-p)ln[ln(\tau_T)]/\tau_T$ (HQIC-type penalty).[10] Andrews (2000) reports simulation evidence that the BIC-type penalty works best in his context, and this has been our experience with $RMSC$. It is desirable to establish an optimal choice for the penalty term, but this is a non-trivial issue that is left for future research.

A simulation study was undertaken to investigate the finite sample properties of our method in this type of setting. We used the following data generating process:

$$y_t = \theta_0 x_t + u_t, \tag{8}$$

$$x_t = \pi' z_t + v_t, \tag{9}$$

where $\theta_0 = 0$, $\pi = [\gamma, 0_{1\times 11}]'$, $0_{1\times 11}$ is a 1×11 vector of zeros, $[u_t, v_t, z_t']' \sim NID(0,\Sigma)$ and Σ is a matrix whose diagonal elements are all equal to one and whose only non-zero off diagonal elements are the $(1,2)$ and $(2,1)$ entries which are both equal to 0.5. We use $\gamma = 1$ and $\gamma = 1/3$ so that the first stage R^2 in population are 0.5 and 0.1, respectively. The candidate set of moment conditions is given by: $E[z_t(y_t - x_t\theta_0)] = 0$. Since the only difference between elements of the candidate set derives from the instrument vector, we index z_t by c so that $f(v_t,\theta;c) = z_t(c)(y_t - x_t\theta)$. Notice that within this design the relevant

[10] See Hannan and Quinn (1979).

moment condition is the one involving the first element of z_t and so $c_r = (1, 0 \ldots 0)'$. Within this framework, the Two-Stage Least Squares (2SLS) is asymptotically equivalent to the GMM estimator with the optimal weighting matrix and so all results reported below are based on the 2SLS estimator.[11]

To compute our moment selection criterion, we need a consistent estimator of the asymptotic variance-covariance matrix. For the 2SLS estimator, we use

$$\hat{V}_{\theta,T}(c) \;=\; \hat{\sigma}_u^2(c) \left[\frac{1}{T} \sum_{t=1}^{T} x_t z_t(c)' \left(\frac{1}{T} \sum_{t=1}^{T} z_t(c) z_t(c)' \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} z_t(c) x_t \right]^{-1} \tag{10}$$

where $\hat{\sigma}_u^2(c) \;=\; (1/T) \sum_{t=1}^{T} \hat{u}_t^2(c)$, $\hat{u}_t(c) = y_t - \hat{\theta}_T(c) x_t$ and $\hat{\theta}_T(c)$ is the 2SLS estimator of θ_0 based on $E[z_t(c)(y_t - x_t\theta)] = 0$. The moment selection procedure is implemented with the penalty term associated with the BIC-type criterion, and so

$$RMSC(c) \;=\; \ln[|\hat{V}_{\theta,T}(c)|] \;+\; \frac{(c'c - 1)\ln(T^{1/2})}{T^{1/2}}. \tag{11}$$

Theorem 3 is premised on the assumption that θ_0 is identified by all the subsets of candidate set considered. This would not be the case if RMSC is minimized over C here because identification rests crucially on $z_{t,1}$, the first element of z_t. We therefore consider the case where RMSC is minimized over the following 12 choices of c: $c = [1'_{q\times 1} 0'_{(12-q)\times 1}]'$, $q = 1, 2, ..., 12$ and $0_{b\times 1}$ is the $b \times 1$ null vector.[12] We define $\hat{q} = |\hat{c}_T|$. The number of Monte Carlo replications is set to 10000, and the sample sizes used are $T = 100$ and $T = 500$.

Tables 1 and 2 report the median bias of the 2SLS estimator and the coverage probabilities of the 90 percent confidence intervals of the 2SLS estimator, respectively, for all twelve choices of c considered and also $\hat{c}_T$. It can be seen that the median bias and the coverage probability tend to deteriorate as the number of redundant instruments increases. However, the use of RMSC leads to a considerable improvement in the quality of the asymptotic approximation - particularly compared to the brute force case in which all twelve

[11] Recall 2SLS is the GMM estimator based on $E[z_t u_t] = 0$ with weighting matrix $W_T = (T^{-1} Z'Z)^{-1}$.

[12] However, see Sections 5 and 6.

elements of the candidate set are used. Lastly, Table 3 shows the summary statistics of the selected number of instruments. The results confirm our asymptotic consistency result in that the number of instruments tends to converge to one as the sample size increases.

As noted above, Theorem 3 is premised on the assumption that θ_0 is identified by all subsets of the candidate set considered. This clearly may not be the case. To understand how the method behaves in such circumstances, it is first useful to explore the connections between redundancy and weak identification. This is done in the next section, and then, using the insights gained from this discussion, we examine in Section 5 the limiting behaviour of RMSC the parameter vector may be weakly identified for some or all of the combinations considered.

4 Redundancy and Weak Identification

The concepts of redundancy and weak identification were introduced to describe superficially very different information scenarios. However, both involve situations in which a set of moments provides marginal information at best, and so in this section we examine whether there is in fact a closer connection between the two concepts. Since weak identification involves a Pitman drift and redundancy does not, the concepts are not easily compared in their original form. Therefore, we introduce a generalization of redundancy that involves a Pitman drift that we refer to as *near redundancy* and explore its properties as part of the discussion. As it will be seen, this extension enables us to delineate the circumstances under which moments that are (nearly) redundant in one setting, provide weak identification in another. For expositional brevity, we frame all this discussion in the context of a linear model estimated via instrumental variables (IV).[13]

In spite of the comments above, it is useful for the purposes of comparison to begin by

[13]The concept of near-redundancy is not specific to linear models. However we restrict attention to this framework here for conformity with the analysis in the following section that is confined to the linear static model.

briefly describing the condition for redundancy within the linear IV setting. To this end, we consider the model:

$$y_t = x_t^{'}\theta_0 + u_t, \tag{12}$$

$$x_t = \Pi_1 z_{1,t} + \Pi_2 z_{2,t} + e_t, \tag{13}$$

where y_t is a scalar, x_t is $p \times 1$ vector, $z_{i,t}$ is $q_i \times 1$ for $i = 1, 2$. In addition, we set $z_t = (z_{1,t}^{'}, z_{2,t}^{'})'$ and set $q = q_1 + q_2$. Let Z_i be the $T \times q_i$ matrix whose t^{th} row is $z_{i,t}^{'}$, $Z = (Z_1, Z_2)$, and X be the $T \times p$ matrix whose t^{th} row is $x_t^{'}$, and u be the $T \times 1$ vector with t^{th} element u_t. We define $v_t = (x_t^{'}, z_t^{'}, u_t, e_t^{'})'$ and assume $\{v_t;\, t = 1, 2, \ldots T\}$ is an i.i.d. sequence of random vectors. Furthermore, it is assumed that $E[u_t|z_t] = 0$ and $E[u_t^2|z_t] = \sigma_0^2$.

Breusch, Qian, Schmidt, and Wyhowski (1999) show that the condition for redundancy of $E[z_{2,t}u_t] = 0$ for the estimation of θ_0 given $E[z_{1,t}u_t] = 0$ is

$$E[z_{2,t}x_t^{'}] \; - \; E[z_{2,t}z_{1,t}^{'}](E[z_{1,t}z_{1,t}^{'}])^{-1}E[z_{1,t}x_t^{'}] \; = \; 0 \tag{14}$$

It can be verified that this condition is equivalent to the restriction that $\Pi_2 = 0$ in (13). For our purposes here, one particular aspect of this definition is worth noting. The moment condition in (14) holds for every t, and this, of course, implies that the sample analog of this condition holds in the limit, that is

$$T^{-1}Z_2'X \; - \; T^{-1}Z_2'Z_1(T^{-1}Z_1'Z_1)^{-1}T^{-1}Z_1'X \; \xrightarrow{p} \; 0 \tag{15}$$

However, since redundancy is a statement about limiting behaviour, it is the condition in (15) that is really important.

To introduce the concept of near redundancy within the linear model, it is necessary to modify the data generation process. It is assumed that (12) still holds but the reduced form is now:

$$x_t \; = \; \Pi_1 z_{1,t} \; + \; \Pi_{2,T} z_{2,t} \; + \; e_t, \tag{16}$$

The key difference is that the coefficient on $z_{2,t}$ depends on T. This means that x_t, and, consequently, v_t depend on T. However, for simplicity, we suppress this dependence in the

notation except at places where it is needed for emphasis. The distributions of all other variables are assumed to be independent of T. Define $E[T^{-1}Z_i'Z_j] = \Omega_{i,j}$, for $i,j = 1,2$, $E[T^{-1}Z_i'X] = \Omega_{i,x,T}$, $\Omega_{i,x} = \lim_{T\to\infty}\Omega_{i,x,T}$, and finally let

$$\Omega_{z,z} \;=\; \begin{bmatrix} \Omega_{1,1} & \Omega_{1,2} \\ \Omega_{2,1} & \Omega_{2,2} \end{bmatrix}, \qquad \Omega_{z,x} \;=\; \begin{bmatrix} \Omega_{1,x} \\ \Omega_{2,x} \end{bmatrix}.$$

We impose the following high level assumptions.

Assumption 5 *(i) rank$\{\Omega_{i,i}\} = q_i$ for $i = 1,2$; (ii) $T^{-1}Z'Z \xrightarrow{p} \Omega_{z,z}$; (iii) $T^{-1/2}Z'u \xrightarrow{d} N(0,\sigma_0^2\Omega_{z,z})$; (iv) $T^{-1/2}\sum_{t=1}^{T} z_t \otimes e_t \xrightarrow{d} N(0,\Sigma_1)$; (v) $T^{-1}u'u \xrightarrow{p} \sigma_0^2$; (vi) $T^{-1/2}\sum_{t=1}^{T}(e_tu_t - \sigma_{e,u}) \xrightarrow{d} N(0,\Sigma_2)$ where $\sigma_{e,u} = E[e_tu_t]$.*

Within this model, we define near redundancy as follows.

Definition 3 Near Redundancy

Let the data be generated via (12) and (16) and Assumption 5 hold. The moment condition $E[z_{2,t}u_t] = 0$ is said to be nearly redundant for the estimation of θ_0 given $E[z_{1,t}u_t] = 0$ if

$$\Omega_{2,x,T} \;-\; \Omega_{2,1}\Omega_{1,1}^{-1}\Omega_{1,x,T} \;=\; T^{-1/2}\eta \tag{17}$$

where η is a matrix of finite constants.

Notice that (17) implies: $\Omega_{2,x} - \Omega_{2,1}\Omega_{1,1}^{-1}\Omega_{1,x} = 0$; and so (15) holds. Therefore, it would be anticipated that nearly redundant moment conditions make no contribution to the limiting variance of the estimator. This intuition is confirmed in the following result.

Theorem 4 *Let $\hat{\theta}_T$ be the 2SLS estimator of θ_0 based on $E[z_tu_t] = 0$. Assume that $E[z_{2,t}u_t] = 0$ is nearly redundant for θ_0 given $E[z_{1,t}u_t] = 0$. Let Assumption 5 hold, the data be generated via (12) and (16), and rank$(\Omega_{1,x}) = p$. The limiting distribution of this GMM estimator is: $T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(0, \sigma_0^2[\Omega_{1,x}'\Omega_{1,1}^{-1}\Omega_{1,x}]^{-1}\right)$.*

For the comparison with weak identification, it is useful to establish the parametric restriction within (12) and (16) that yields near redundancy of $E[z_{2,t}u_t] = 0$. By definition, we have

$$\Omega_{2,x,T} - \Omega_{2,1}\Omega_{1,1}^{-1}\Omega_{1,x,T} \;=\; E[T^{-1}Z_2'X] - E[T^{-1}Z_2'Z_1]\left\{E[T^{-1}Z_1'Z_1]\right\}^{-1} E[T^{-1}Z_1'X] \quad (18)$$

Using (16) and Assumption 5, it follows from (18) that

$$\Omega_{2,x,T} \;-\; \Omega_{2,1}\Omega_{1,1}^{-1}\Omega_{1,x,T} \;=\; (\Omega_{2,2} \;-\; \Omega_{2,1}\Omega_{1,1}^{-1}\Omega_{1,2})\Pi_{2,T}' \quad (19)$$

Assumption 5(i) implies $(\Omega_{2,2} - \Omega_{2,1}\Omega_{1,1}^{-1}\Omega_{1,2})$ is a nonsingular matrix of constants and so, taken together, (17) and (19) imply that $E[z_{2,t}u_t] = 0$ is nearly redundant given $E[z_{1,t}u_t] = 0$ if and only if $\Pi_{2,T} = C_2T^{-1/2}$ for some matrix of constants C_2.

We now consider weak identification. For our purposes, it suffices to consider the "classic" version of weak identification in the linear model that is analyzed by Staiger and Stock (1997).[14] So we assume the data is generated by (12) and (16) with $\Pi_1 = 0$, $\Pi_{2,T} = T^{-1/2}C_2$ and estimation is based on $E[z_{2,t}u_t] = 0$. In this case, the key derivative matrix is

$$\Omega_{2,x,T} \;=\; E[T^{-1}Z_2'X] \;=\; T^{-1/2}\Omega_{2,2}C_2'$$

and so $\Omega_{2,x} = 0_{q_2\times p}$ causing identification to fail.

Since $\Pi_{2,T}$ behaves the same way under near redundancy and weak identification, it is natural to wonder whether a set of moment conditions can be nearly redundant when other moments are included but be associated with weak identification if these other moments are excluded. To explore this question, we consider the case in which the data are generated by (12) and (16) with $rank\{\Pi_1\} = p$, $\Pi_{2,T} = T^{-1/2}C_2$. If estimation is based on $E[z_tu_t] = 0$ then $E[z_{2,t}u_t] = 0$ is nearly redundant because of the inclusion of $E[z_{1,t}u_t] = 0$. However, if estimation is based on $E[z_{2,t}u_t] = 0$ alone then this set of moments is inevitably the only

[14] See Zivot, Startz, and Nelson (2003) for a discussion of different scenarios that can lead to weak identification in this model.

source of information about θ_0. Does this mean that θ_0 is weakly identified? The answer is may be or may be not. To see this, note that within this specification

$$\Omega_{2,x,T} \;=\; \Omega_{2,1}\Pi_1' \;+\; T^{-1/2}\Omega_{2,2}C_2'$$

Therefore, θ_0 is identified provided $rank\{\Omega_{2,1}\Pi_1^{'}\} = p$, but weakly identified if this condition fails. Or, put another way, θ_0 is identified provided $z_{2,t}$ inherits enough of the explanatory power of $z_{1,t}$ for x_t when the latter is omitted. However, notice that if θ_0 is weakly identified based on $E[z_{2,t}u_t] = 0$ alone then these moments are nearly redundant once the moment condition is augmented by $E[z_{1,t}u_t] = 0$.

5 Analysis of RMSC when Weak Identification is a Possibility

In this section, we continue our analysis of RMSC. As noted above, the consistency result in Theorem 3 is derived under certain regularity conditions. One of these conditions is the requirement that θ_0 is identified by all the subsets of the candidate set over which the minimization is performed. It is clear from the discussion in Section 4 that this is a viable scenario. However, it is also clear that moment conditions that are redundant given the relevant subset may fail to identify θ_0 when some or all of that relevant subset are excluded from the estimation. Therefore, in this section, we consider the limiting behaviour of RMSC in the presence of weak identification, and then use these results to derive the limiting behaviour of $\hat{c}_T$ when the parameter vector may be weakly identified for some of the combinations considered.

The analysis is undertaken in the context of the linear model in (12) and (16). However, this time, we partition to x_t into $(x_{1,t}^{'}, x_{2,t}^{'})'$ where $x_{i,t}$ is $p_i \times 1$ for $i = 1, 2$, and partition θ_0 conformably into $(\theta_{0,1}^{'}, \theta_{0,2}^{'})'$. For what follows, it is useful to take account of this partition

in the presentation of the data generation process for x_t. Therefore, we rewrite (16) as

$$x_{i,t} \;=\; \Pi_{i,1,T} z_{1,t} \;+\; \Pi_{i,2,T} z_{2,t} \;+\; e_{i,t}, \qquad \text{for } i = 1, 2 \tag{20}$$

where $\Pi_{i,j,T}$ is $p_i \times q_j$. All other definitions are as above; we also impose Assumption 5 as before. Once again, let $\hat{\theta}_T$ be the 2SLS estimator of θ_0 based on $E[z_t u_t] = 0$. From Assumption 5, it follows that the variance of the limiting distribution of 2SLS is $V_\theta = \sigma_0^2(\Omega_{x,z}\Omega_{z,z}^{-1}\Omega_{z,x})^{-1}$. Given this structure, the obvious candidate for the covariance matrix estimator is: $\hat{V}_{\theta,T} = \hat{\sigma}_T^2[T^{-1}X'Z(T^{-1}Z'Z)^{-1}T^{-1}Z'X]^{-1}$; where $\hat{\sigma}_T^2 = T^{-1}(y - X\hat{\theta}_T)'(y - X\hat{\theta}_T)$, and so

$$ln[|\hat{V}_{\theta,T}|] \;=\; pln(\hat{\sigma}_T^2) \;-\; ln\left[|T^{-1}X'Z(T^{-1}Z'Z)^{-1}T^{-1}Z'X|\right]. \tag{21}$$

As remarked above, the aim of this section is to analyze the behaviour of RMSC when θ_0 is only weakly identified by some subsets of the candidate set of moment conditions. To achieve this end, it is useful to consider first the behaviour of $ln[|\hat{V}_{\theta,T}|]$ for three distinct scenarios. These are as follows:

- *Scenario I: θ_0 is weakly identified*

  $\Pi_{i,j,T} = T^{-1/2}C_{i,j}$ for some matrices of constants $C_{i,j}$, $i, j = 1, 2$ and $rank\{C\} = p$ where

  $$C \;=\; \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix}.$$

- *Scenario II: $\theta_{0,1}$ is identified but $\theta_{0,2}$ is weakly identified*

  $\Pi_{1,1,T} = \Pi_{1,1}$ with $rank(\Pi_{1,1}) = p_1$; $\Pi_{1,2,T} = T^{-1/2}C_{1,2}$, for some matrix of constants $C_{1,2}$; $\Pi_{2,j,T} = T^{-1/2}C_{2,j}$ for some matrices of constants $C_{2,j}$, $j = 1, 2$; $rank\{[C_{2,1}, C_{2,2}]\} = p_2$.

- *Scenario III: θ_0 is identified*

  $\Pi_{i,1,T} = \Pi_{i,1} + T^{-1/2}C_{i,1}$ with $rank(\Pi_{i,1}) = p_i$ for $i = 1, 2$; $\Pi_{i,2,T} = T^{-1/2}C_{i,2}$ for some matrices of constants $C_{i,j}$, $i, j = 1, 2$.

Two aspects of Scenario III are worthy of comment. First, note the specification for $\Pi_{i,1,T}$ implies that θ_0 is identified by $E[z_{1,t}u_t] = 0$ but allows identification to rest on different elements of the moment condition for $\theta_{0,1}$ and $\theta_{0,2}$. Second, some elements of $E[z_t u_t] = 0$ are nearly redundant given other elements.[15]

The following theorem presents the large sample behaviour of $ln[|\hat{V}_{\theta,T}|]$ under these three scenarios.

Theorem 5 *Let the data be generated via (12) and (20) and Assumption 5 hold.*

(i) Under Scenario I: $ln[|\hat{V}_{\theta,T}|] \; = \; pln(T) \; + \; O_p(1)$*;*

(ii) Under Scenario II: $ln[|\hat{V}_{\theta,T}|] \; = \; p_2 ln(T) \; + \; O_p(1)$*;*

(iii) Under Scenario III: $ln[|\hat{V}_{\theta,T}|] \; = \; pln[\sigma_0^2] \; - \; ln[|\Omega_{1,x}' \Omega_{1,1}^{-1} \Omega_{1,x}|] \; + \; o_p(1) \; = \; O_p(1)$

We now consider the implications of Theorem 5 for moment selection based on RMSC in which the equation of interest is (12), the candidate set of moments is given by $E[z_t(y_t - x_t'\theta_0)] = 0$ and the relationship between x_t and z_t is given by (20). Using the notation from Section 3 and specializing the definition of RMSC to the model in this section, the chosen selection vector is[16]

$$\hat{c}_T \; = \; min_{c \in C} \left\{ ln[\hat{\sigma}_T^2(c)] \; - \; ln\left[\left| T^{-1} X' Z(c) \{ T^{-1} Z(c)' Z(c) \}^{-1} T^{-1} Z(c)' X \right| \right] \; + \; \kappa(|c|, T) \right\}$$

where $Z(c)$ is the $T \times |c|$ matrix whose t^{th} row is $z_t(c)'$, $\hat{\sigma}_T^2(c) = T^{-1}[y - X\hat{\theta}_T(c)]'[y - X\hat{\theta}_T(c)]$ and $\hat{\theta}_T(c)$ is the 2SLS estimator of θ_0 based on $E[z_t(c)u_t] = 0$.

We now require two additional assumptions that specify respectively a partition of C and an identification condition.

[15] $E[z_{2,t}u_t] = 0$ is nearly redundant given $E[z_{1,t}u_t] = 0$. Elements of $E[z_{1,t}u_t] = 0$ may also be nearly redundant given the remaining elements of this vector depending on the elements of $\Pi_{i,1}$ and $C_{i,1}$.

[16] Since we consider behaviour over different choices of instrument, we now reinstate the indexing by c.

Assumption 6 *$C = C_I \cup C_{II} \cup C_{III}$ where C_I yields models that fit within Scenario I, C_{II} yields models that fit within Scenario II and C_{III} yields models that fit within Scenario III.*

To facilitate the presentation of the identification condition, we define $V_\theta(c)$ to be the variance of the limiting distribution of the 2SLS estimator based on $E[z_t(c)u_t] = 0$. Note that within the framework here, the minimum value for $V_\theta(c)$ is $V(\iota_q)$ where ι_q is $q \times 1$ vector of ones.

Assumption 7 *There is a $c_r \in C_{III}$ that satisfies the properties in Definition 2 with $q_{max} = q$ and $\mathcal{C}_{min} = \{c_r\}$.*

The following theorem gives the limiting behaviour of RMSC when subsets of the candidate set provide only weak identification.

Theorem 6 *Let the data be generated by (12) and (20) and Assumptions 4–7 hold, then $\hat{c}_T \xrightarrow{p} c_r$.*

Theorem 6 indicates that RMSC is consistent for c_r in this model even when some subsets of the candidate set provide only weak identification.

To conclude this section, we explore the finite sample behaviour of $RMSC$ in a setting where weak identification is a possibility. Simulated data are generated from the model in (12)-(13) with $p = 1$, $q_1 = 2$ and $q_2 = 6$. The primitive variables are $v_t' = [u_t, e_t, z_t']$, and random samples are generated under the assumption that $v_t \sim N(0, \Sigma_v)$ where the main diagonal elements of Σ_v are all set to unity, and the only non-zero off diagonal elements are $cov(u_t, e_t) = \sigma_{ue}$, that is, $\Sigma_v(1,2)$ and $\Sigma_v(2,1)$. In all experiments, $\theta_0 = 0.1$, $\Pi_2 = 0$ and the elements of Π_1 are given by $\pi_{1,1} = 0.6243k$ and $\pi_{1,2} = 0.3660k$ where k is chosen so that $\Pi_1'\Pi_1 = R_f^2/(1-R_f^2)$ for some fixed value of R_f^2, the multiple correlation coefficient of the reduced form equation, (13).[17] Each experiment consists of a specification of (T, R_f^2, σ_{ue}) from the following sets: $T \in \{100, 500\}$; $R_f^2 \in \{0.1, 0.5\}$; $\sigma_{ue} \in \{0.1, 0.5, 0.9\}$.

[17] This design exploits results presented in Hahn and Inoue (2002).

Three aspects of this design are worth noting. First, $E[z_{2,t}u_t] = 0$ is redundant for θ_0 given $E[z_{1,t}u_t] = 0$, and so $c_r = (1, 1, 0_{1\times 6})'$. Second, θ_0 is unidentified by $E[z_t(c)u_t] = 0$ for any $c = (0, 0, n')'$. Third, previous research suggests that $R_f^2 = 0.1$ may be associated with weak identification problems in small to moderate sized samples.[18]

For each replication, the 2SLS estimator is calculated based on $E[z_t(c)u_t] = 0$ for all $c \in C$ and $\hat{c}_T$ is calculated using the BIC-type penalty; see Comment 8. Table 4 reports a summary statistics associated with the distribution of $\hat{c}_T$ and the post-selection estimator $\hat{\theta}_T(\hat{c}_T)$. Since there are 255 possible combinations of instruments, we group these possibilities into six cases: $1R, 2R, 1R/I, 2R/I^*, I$ and *all*, where $1R$ denotes the cases in which $c = (a, 0_6')'$ for $a \in \{(1, 0), (0, 1)\}$, implying that that the selection vector consists of only one of the relevant instruments; $2R$ denotes the case in which $c = (1, 1, 0_6')'$, indicating that the selection vector consists of only both relevant instruments; $1R/I$ denotes the cases in which $c = (a', b')'$ for a given above and $b \neq 0_6$, meaning that the selection vector consists of one relevant instrument and at least one redundant instrument; $2R/I^*$ denotes the cases in which $c = (1, 1, d')'$ and $d \neq 0_6$ *or* ι_6, that is, the selection vector consists of both relevant and at least one but not all six redundant instruments; I denotes the cases in which $c = (0, 0, b')'$ for b given above, implying that the selection vector consists of only redundant instruments; and finally, *all* denotes the case in which $c = \iota_8'$, indicating that the selection vector contains all eight instruments, that is, the two relevant instruments as well as the six redundant instruments.

Table 4 reports the results for R_f^2 equal to 0.1 and 0.5. With $R_f^2 = 0.1$, it can be seen that for $T = 100$, RMSC tends to pick combinations that include one or more of the relevant instruments but also tends to include at least some of the irrelevant instruments as well. However, by $T = 500$, the method is clearly doing a better job of identifying the relevant set. This improvement is also reflected in the coverage probabilities of the 90% confidence intervals based on the limiting distribution: for $T = 100$, the actual coverage rate is clearly

[18] For example, see Hahn and Inoue (2002).

different from the nominal level, but by $T = 500$, it is very close to the nominal value except when the regressor is highly endogenous. We conjecture that these distortions are a further manifestation of the problems caused by weak identification.[19] With $R_f^2 = 0.5$, everything works much better. RMSC identifies the relevant instruments with high probability, and the coverage probabilities are close to the nominal level at both sample sizes regardless of the value of σ_{ue}.

6 Concluding Remarks

In this paper, we make five contributions to the literature on information and entropy in Generalized Method of Moments (GMM) estimation. First, we introduce the concept of the long run canonical correlations (LRCC's) between the true score vector and the moment function $f(v_t, \theta_0)$ and show that they provide a metric for the information contained in the population moment condition $E[f(v_t, \theta_0)] = 0$. Second, we show that the entropy of the limiting distribution of the GMM estimator can be written in terms of these LRCC's. Third, motivated by the aforementioned results, we introduce an information criterion based on this entropy that can be used as a basis for moment selection. Fourth, we introduce the concept of nearly redundant moment conditions and use it to explore the connection between redundancy and weak identification. Fifth, we analyze the behaviour of the aforementioned entropy based moment selection method in two scenarios of interest; these scenarios are: (i) nonlinear dynamic models where the parameter vector is identified by all the combinations of moment conditions considered; (ii) linear static models where the parameter vector may be weakly identified for some of the combinations considered.

[19] For example, see Nelson and Startz (1990) or Hall, Rudebusch, and Wilcox (1996).

Appendix:

Long Run Canonical Correlations (LRCC's):

Definition A.1 *Let x_t and z_t be $p \times 1$ and $q \times 1$ and $m = min(p, q)$. Suppose that $T^{-1/2} \sum_{t=1}^{T} v_t \xrightarrow{d} N(0, \Sigma_v)$ where $v_t = (x_t', z_t')'$, $\Sigma_v = \lim_{T\to\infty} Var[T^{-1/2} \sum_{t=1}^{T} v_t]$ is a finite positive definite matrix and*

$$\Sigma_v = \begin{bmatrix} \Sigma_{x,x} & \Sigma_{x,z} \\ \Sigma_{z,x} & \Sigma_{z,z} \end{bmatrix}$$

using the obvious notation. The population LRCC's between x_t and z_t are denoted by $\{\rho_i; i = 1, 2, \ldots m\}$, where by convention $\rho_i \geq 0$ for $i = 1, 2, \ldots m$ and $\rho_i \geq \rho_{i+1}$ for $i = 1, 2, \ldots m-1$, and have the following properties: (i) $\{\rho_i^2\}$ are the m largest solutions to the determinantal equation $|\Sigma_{x,z}\Sigma_{z,z}^{-1}\Sigma_{z,x} - \rho^2\Sigma_{x,x}| = 0$; (ii) $\rho_i = a_i'\Sigma_{x,z}b_i$ where a_i and b_i satisfy $(\Sigma_{x,z}\Sigma_{z,z}^{-1}\Sigma_{z,x} - \rho_i^2\Sigma_{x,x})a_i = 0$ and $(\Sigma_{z,x}\Sigma_{x,x}^{-1}\Sigma_{x,z} - \rho_i^2\Sigma_{z,z})b_i = 0$ for $i = 1, 2, \ldots m$.[20]

Using similar arguments to Rao (1973)[p.583], Jana (2005) shows that the following properties hold.

Lemma A.1 *Let $m = p$ (in Definition A.1) and A be the $p \times p$ matrix with i^{th} column a_i. Then, the following identities hold: $\Sigma_{x,x} = A'^{-1}A^{-1}$; $\Sigma_{x,z}\Sigma_{z,z}^{-1}\Sigma_{z,x} = A'^{-1}R^2A^{-1}$; where $R = diag(\rho_1, \rho_2, \ldots \rho_p)$.*

Proof of Theorem 1(a):

To simplify the presentation, we set $f_t(\theta) = f(v_t, \theta)$. Assumption 3(iii) implies that

$$\int_{\mathcal{V}_{(-\infty,t)}} f_t(\theta)p(V_t|\theta)dV_t = 0 \tag{22}$$

for $\theta \in \mathcal{N}_{\theta_0}$. Substituting $p(V_t|\theta) = p(v_t|V_{t-1}, \theta)p(V_{t-1}, \theta)$ into (22) and differentiating

[20] Recall that the linear combinations are chosen so as to normalize the variances to one, that is $a_i'\Sigma_{x,x}a_i = b_i'\Sigma_{z,z}b_i = 1$.

under the integral sign, it follows from the Lebesgue dominated convergence theorem that

$$0 = \int_{\mathcal{V}_{(-\infty,t)}} \frac{\partial f_t(\theta)}{\partial\theta}\Big|_{\theta=\theta_0} p(V_t|\theta_0)dV_t + \int_{\mathcal{V}_{(-\infty,t)}} f_t(\theta_0)\left[\frac{\partial p(v_t|V_{t-1},\theta)}{\partial\theta}\right]'\Big|_{\theta=\theta_0} p(V_{t-1}|\theta_0)dV_t$$
$$+ \int_{\mathcal{V}_{(-\infty,t)}} f_t(\theta_0)p(v_t|V_{t-1},\theta_0)\left[\frac{\partial p(V_{t-1}|\theta)}{\partial\theta}\right]'\Big|_{\theta=\theta_0} dV_t\,. \quad (23)$$

Since

$$s_t(\theta) = [\partial p(v_t|V_{t-1},\theta)/\partial\theta][1/p(v_t|V_{t-1},\theta)]\,, \quad (24)$$

equation (23) can be rewritten as

$$0 = G(f) + E[f_t(\theta_0)s_t(\theta_0)'] + \int_{\mathcal{V}_{(-\infty,t)}} f_t(\theta_0)p(v_t|V_{t-1},\theta_0)\left[\frac{\partial p(V_{t-1}|\theta)}{\partial\theta}\right]'\Big|_{\theta=\theta_0} dV_t\,. \quad (25)$$

Now $p(V_{t-1}|\theta_0) = \{\prod_{n=1}^{N} p(v_{t-n}|V_{t-n-1},\theta_0)\}p(V_{t-N-1}|\theta_0)$, and so it follows that

$$\frac{\partial p(V_{t-1}|\theta)}{\partial\theta}\Big|_{\theta=\theta_0} = \sum_{n=1}^{N} \frac{\partial p(v_{t-n}|V_{t-n-1},\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\left[\frac{1}{p(v_{t-n}|V_{t-n-1},\theta_0)}\right] p(V_{t-1}|\theta_0)$$
$$+ \prod_{n=1}^{N} p(v_{t-n}|V_{t-n-1},\theta_0)\,\frac{\partial p(V_{t-N-1}|\theta)}{\partial\theta}\Big|_{\theta=\theta_0} \quad (26)$$

Using (24) and (26) in (25), we obtain

$$0 = G(f) + \sum_{n=0}^{N} E[f_t(\theta_0)s_{t-n}(\theta_0)']$$
$$+ \int_{\mathcal{V}_{(-\infty,t)}} f_t(\theta_0)\prod_{n=0}^{N} p(v_{t-n}|V_{t-n-1},\theta_0)\left[\frac{\partial p(V_{t-N-1}|\theta)}{\partial\theta}\right]'\Big|_{\theta=\theta_0} dV_t$$
$$= G(f) + \sum_{n=0}^{N} E[f_t(\theta_0)s_{t-n}(\theta_0)'] + E\left\{f(v_t,\theta_0)\left[\frac{\partial \ln p(V_{t-N-1}|\theta)}{\partial\theta}\right]_{\theta=\theta_0}\right\}. \quad (27)$$

Since the second term in (27) is bounded by $O(\sum_{n=0}^{N}\alpha_n^{1-1/\zeta-1/\eta}) = O(1)$ and the third term is of order $O(\alpha_{N+1}^{1-1/\zeta-1/\eta})$ by the mixing inequality, taking the limit as $N\to\infty$ in (27) gives the desired result.

Proof of Theorem 1(b):

By definition,

$$C(f) = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\sum_{k=1}^{T} E[f(v_t,\theta_0)s_k(\theta_0)'] \quad (28)$$

The conditional score vector, $s_t(\theta_0)$ is a martingale difference sequence with respect to Ω_{t-1}, the σ–algebra generated by V_{t-1}. Therefore, $E[f(v_t,\theta_0)s_k(\theta_0)^{'}] = 0$ for all $k > t$. Using this result and Assumption 1 in (28), $C(f)$ becomes,

$$\begin{aligned} C(f) &= \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\sum_{n=t}^{T} E[f(v_n,\theta_0)s_t(\theta_0)^{'}] = \lim_{N\to\infty}\sum_{n=0}^{N-1}\left(1-\frac{n}{N}\right)E[f(v_t,\theta_0)s_{t-n}(\theta_0)^{'}] \\ &= \lim_{N\to\infty}\sum_{n=0}^{N-1} E[f(v_t,\theta_0)s_{t-n}(\theta_0)^{'}] \end{aligned}$$

where the last equality follows from the mixing inequality and the assumption that $\sum_{j=0}^{\infty} j\alpha_j^{1-1/\zeta-1/\eta} < \infty$. The desired result then follows from Theorem 1(a).

Proof of Theorem 2

Part (i): From the definition of $\{\rho_i(f);\, i = 1,2,\dots p\}$ and Theorem 1(b) it follows that $\{\rho_i(f);\, i = 1,2,\dots p\}$ are the eigenvalues of $\mathcal{I}_\theta^{-1/2}G(f)^{'}S(f)^{-1}G(f)\mathcal{I}_\theta^{-1/2}$. By construction $\mathcal{I}_\theta^{-1/2}G(f)^{'}S(f)^{-1}G(f)\mathcal{I}_\theta^{-1/2}$ is symmetric and $rank\{\mathcal{I}_\theta^{-1/2}G(f)^{'}S(f)^{-1}G(f)\mathcal{I}_\theta^{-1/2}\} = rank(G(f))$ because both $S(f)^{-1}$ and $\mathcal{I}_\theta^{-1/2}$ are nonsingular. The result then follows directly.

Part (ii): Using the martingale difference property of the score vector, it can be shown that

$$\lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} s_t(\theta_0)] \;=\; \mathcal{I}_\theta$$

From Theorem 1(b), it follows that $V_\theta(f)^{-1} = C(f)'S(f)^{-1}C(f)$. Therefore, the population LRCC's between $f(v_t,\theta_0)$ and $s_t(\theta_0)$ are the solutions to $|V_\theta^{-1} - \rho^2\mathcal{I}_\theta| = 0$. The result then follows from Lemma A.1.

Part (iii): Part (a) follows trivially from $V_\theta(f_1) = V_\theta(f_2)$ and the definition of the LRCC. Now consider part (b). From Dhrymes (1984)[Proposition 65], $V_\theta(f_1) - V_\theta(f_2)$ is positive semi-definite (psd) if and only if $[V_\theta(f_2)]^{-1} - [V_\theta(f_1)]^{-1}$ is psd. Set $M = [V_\theta(f_2)]^{-1} - [V_\theta(f_1)]^{-1}$. From Assumption 2, it follows that

$$M \;=\; G(f_2)'S(f_2)^{-1}G(f_2) \;-\; G(f_1)'S(f_1)^{-1}G(f_1) \tag{29}$$

Since $\mathcal{I}_\theta$ is a symmetric positive definite matrix, there exists nonsingular matrix N such that $\mathcal{I}_\theta = NN'$. Given the properties of N it follows that M is psd if and only if

$N^{-1}M(N^{-1})'$ is positive semi-definite. By definition, $\{\rho_i^2(f_j); i = 1, 2, \ldots p\}$ are the eigenvalues of $N^{-1}G(f_j)'S(f_j)^{-1}G(f_j)(N^{-1})'$. Therefore, it follows from Magnus and Neudecker (1991)[Theorem 9, p.208] that if M is psd then

$$\rho_i^2(f_2) \geq \rho_i^2(f_1), \text{ for } i = 1, 2, \ldots p \tag{30}$$

To establish that the inequality is strict for at least one i, we consider $trace(N^{-1}M(N^{-1})')$. From the definition of the LRCC, it follows that

$$trace\{N^{-1}M(N^{-1})'\} \; = \; \sum_{i=1}^{p} \rho_i^2(f_2) \; - \; \sum_{i=1}^{p} \rho_i^2(f_1)$$

If $N^{-1}M(N^{-1})'$ is positive semi-definite then $trace\{N^{-1}M(N^{-1})'\} > 0$ and so, using (30) it must follow that $\rho_i^2(f_2) > \rho_i^2(f_1)$ for at least one i. This proves the "if"part; the "only if" is easily deduced by reversing the sequence of the logic and so is omitted for brevity.

Proof of Theorem 3:

We define: $\Delta_T(c, c_r) = RMSC(c) - RMSC(c_r)$. From Definition 2(ii), we have that $V_\theta(\iota_{q_{max}}) = V_\theta(c_r)$ and so it suffices to consider $\Delta_T(c, c_r)$ for two choices of c: (i) c such that $V_\theta(c) = V_\theta(c_r)$; (ii) c such that $V_\theta(c) - V_\theta(c_r) = M(c)$ where $M(c)$ is a non-null psd matrix.

Case (i): c such that $V_\theta(c) = V_\theta(c_r)$

In this case, we have

$$\begin{aligned} \Delta_T(c, c_r) \;\; &= \;\; [\ln|V_{\theta,T}(c)| - \ln|V_\theta(c)|] - [\ln|V_{\theta,T}(c_r)| - \ln|V_\theta(c_r)|] \\ &\quad + \kappa(|c|, T) \; - \; \kappa(|c_r|, T) \end{aligned} \tag{31}$$

By Assumption 4(iii), we have: $\tau_T\Delta_T(c, c_r) = O_p(1) + \tau_T[\kappa(|c|, T) - \kappa(|c_r|, T)]$. Assumption 4(i) states that $|c| > |c_r|$ and so it follows from Assumption 4(iv) that: $\lim_{T\to\infty} \tau_T[\kappa(|c|, T) - \kappa(|c_r|, T)] = +\infty$. Thus, $\tau_T\Delta_T(c, c_r)$ is positive with probability one in the limit as $T \to \infty$.

Case (ii): c such that $V_\theta(c) - V_\theta(c_r) = M(c)$ *where* $M(c)$ *is a non-null positive semi-definite matrix*

From Lemma A.1, it follows that: $\ln|V_\theta(c)| - \ln|V_\theta(c_r)| = m(c, c_r)$; where $m(c, c_r) : C \times C \to [0, +\infty)$. From Assumptions 4(iii) and (iv), it follows that: $\ln|V_{\theta,T}(c)| - \ln|V_{\theta,T}(c_r)| = m(c, c_r) + o_p(1)$. Therefore, $\Delta_T(c, c_r)$ is positive with probability one in the limit as $T \to \infty$. Taken together the results in Cases (i)-(ii) yield the desired result.

Proof of Theorem 4:

From Assumption 5, it follows that $T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, [\Omega'_{z,x}\Omega^{-1}_{z,z}\Omega_{z,x}]^{-1})$. Now consider $V = \Omega_{x,z}\Omega^{-1}_{z,z}\Omega_{z,x}$. Define $V_1 = \Omega'_{1,x}\Omega^{-1}_{z,z}\Omega_{1,x}$. From Hall and Peixe (2003)[equation (19)], it follows that

$$V \;=\; V_1 \;+\; G'FG \tag{32}$$

where $F = (\Omega_{2,2} \;-\; \Omega_{2,1}\Omega^{-1}_{1,1}\Omega_{1,2})^{-1}$ and $G = \Omega_{2,x} - \Omega_{2,1}\Omega^{-1}_{1,1}\Omega_{1,x}$. It follows from (17) that $G = 0$ and so $V = V_1$ which gives the desired result.

Proof of Theorem 5:

Part(i): From Staiger and Stock (1997)[Theorem 1(b)], it follows that $\hat{\sigma}^2_T = O_p(1)$. Now consider $ln[|A_T|]$ where $A_T = B_T D_T B'_T$, $B_T = T^{-1}X'Z$, and $D_T = (T^{-1}Z'Z)^{-1}$. Partition B_T and D_T as follows:

$$B_T \;=\; \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}, \qquad D_T \;=\; \begin{bmatrix} D_{1,1} & D_{1,2} \\ D_{2,1} & D_{2,2} \end{bmatrix} \tag{33}$$

where $B_{i,j}$ is $p_i \times q_j$, $D_{i,j}$ is $q_i \times q_j$ for $i, j = 1, 2$ (and the T subscript on $B_{i,j}$ and $D_{i,j}$ is suppressed for notational simplicity). Using this partition, it follows that

$$A_T \;=\; \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \tag{34}$$

where

$$A_{i,j} \;=\; B_{i,1}D_{1,1}B'_{j,1} \;+\; B_{i,1}D_{1,2}B'_{j,2} \;+\; B_{i,2}D_{2,1}B'_{j,1} \;+\; B_{i,2}D_{2,2}B'_{j,2} \tag{35}$$

and we have suppressed the T subscript on the submatrices of A_T. From Dhrymes (1984)[Proposition 30], it follows that

$$|A_T| \;=\; |A_{2,2}||A_{1,1} \;-\; A_{1,2}A^{-1}_{2,2}A_{2,1}|\,. \tag{36}$$

The order of $|A_T|$ can therefore be deduced from (36) once the orders of $\{A_{i,j}\}$ are known. Assumption 5 implies that $D_{i,j} = O_p(1)$, and Scenario I implies that $B_{i,j} = O_p(T^{-1/2})$. Therefore, $A_{i,j} = O_p(T^{-1})$. Define $\tilde{A}_T = TA_T$, and $\tilde{A}_{i,j} = TA_{i,j}$. Notice that $\tilde{A}_T = O_p(1)$ by construction. We now show that $\tilde{A}_T$ is positive definite with probability one in the limit as $T \to \infty$. To this end, we write $\tilde{A}_T = \tilde{B}_T D_T \tilde{B}_T'$ where $\tilde{B}_T = T^{1/2} B_T$, and consider the following quadratic form: $v'\tilde{A}_T v = \left\{v'(T^{1/2}B_T)\right\} D_T \left\{(T^{1/2}B_T')v\right\} = v'\tilde{B}_T D_T \tilde{B}_T' v$; for some non-zero vector v. Since D_T is positive definite by construction, we need to consider $(T^{1/2}B_T')v = \tilde{B}_T' v$.

In order to do this, we express $x_{1,t}$ and $x_{2,t}$ in the matrix form as follows,

$$X_i = Z_1 \Pi_{i,1,T}' + Z_2 \Pi_{i,2,T}' + E_i, \qquad \text{for } i = 1, 2$$

where X_i and E_i are $T \times p_i$, Z_i is $T \times q_i$, for $i = 1, 2$. Now let $X = [X_1 \; X_2]$ and $Z = [Z_1 \; Z_2]$, then B_T can be written as:

$$B_T = T^{-1} X'Z = \begin{bmatrix} T^{-1}X_1'Z_1 & T^{-1}X_1'Z_2 \\ T^{-1}X_2'Z_1 & T^{-1}X_2'Z_2 \end{bmatrix} = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}.$$

Each block of B_T can be written as

$$B_{i,j} = \Pi_{i,1,T} \left(T^{-1} Z_1' Z_j\right) + \Pi_{i,2,T} \left(T^{-1} Z_2' Z_j\right) + T^{-1} E_i' Z_j$$

for $i, j = 1, 2$. Then the each block of $\tilde{B}_T$, $\tilde{B}_{i,j}$ (supressing the T again), is given by: $\tilde{B}_{i,j} = T^{1/2} B_{i,j}$ for $i, j = 1, 2$.

Using the assumptions of Scenario I, $\Pi_{i,j,T} = T^{1/2} C_{i,j}$, we can conclude that

$$\tilde{B}_T' v \xrightarrow{d} \begin{bmatrix} C_{1,1}\Omega_{1,1} + C_{1,2}\Omega_{2,1} & C_{1,1}\Omega_{1,2} + C_{1,2}\Omega_{2,2} \\ C_{2,1}\Omega_{1,1} + C_{2,2}\Omega_{2,1} & C_{2,1}\Omega_{1,2} + C_{2,2}\Omega_{2,2} \end{bmatrix}' v + \tilde{B}_{normal}' v$$

where $\tilde{B}_{normal}$ is a matrix whose elements are normally distributed. Now let

$$C = \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix}.$$

then we can write, $\tilde{B}_T' v \xrightarrow{d} (C\Omega_{z,z})'v + \tilde{B}_{normal}' v$. By construction, Ω is positive definite. Hence if the matrix C is of full rank (i.e. $rank(C) = p$), $\tilde{A}_T$ is positive definite with probability 1.

Returning to A_T, if we substitute in (36) for $A_{i,j}$ in terms of $\tilde{A}_{i,j}$ then we obtain: $|A_T| = T^{-p}|\tilde{A}_T| = O_p(T^{-p})$, where the last equality follows from the properties of $\tilde{A}_T$ derived above. The desired result then follows from (21).

Part (ii): We first consider $ln[|A_T|]$. The analysis evolves along similar lines to the proof of part (i) and uses the partitions defined therein. Once again, Assumption 5 implies that $D_{i,j} = O_p(1)$. Scenario II implies that $B_{1,j} = O_p(1)$ and $B_{2,j} = O_p(T^{-1/2})$. Therefore, it follows from (35) that $A_{1,1} = O_p(1)$, $A_{1,2} = O_p(T^{-1/2})$, $A_{2,1} = O_p(T^{-1/2})$ and $A_{2,2} = O_p(T^{-1})$. Define

$$\bar{A}_T = \begin{bmatrix} \bar{A}_{1,1} & \bar{A}_{1,2} \\ \bar{A}_{2,1} & \bar{A}_{2,2} \end{bmatrix} \tag{37}$$

where $\bar{A}_{1,1} = A_{1,1}$, $\bar{A}_{1,2} = T^{1/2}A_{1,2}$, $\bar{A}_{2,1} = T^{1/2}A_{2,1}$ and $\bar{A}_{2,2} = TA_{2,2}$.

Now we consider the properties of $\bar{A}_T$. Note $\bar{A}_T = O_p(1)$ by construction. We now show that $\bar{A}_T$ is positive definite with probability one in the limit as $T \to \infty$. To do this, we define,

$$\bar{B}_T = \begin{bmatrix} B_{1,1} & B_{1,2} \\ T^{1/2}B_{2,1} & T^{1/2}B_{2,2} \end{bmatrix}$$

Then, $\bar{A}_T$ can be written as

$$\bar{A}_T = \bar{B}_T D_T \bar{B}_T'$$

By the same logic as in part (i), we need to consider $\bar{B}_T' v$. Each block of $\bar{B}_T$ is:

$$\begin{aligned} \bar{B}_{1,j} &= \Pi_{1,1,T}\left(T^{-1}Z_1'Z_j\right) + \Pi_{1,2,T}\left(T^{-1}Z_2'Z_j\right) + T^{-1}E_1'Z_j \\ \bar{B}_{2,j} &= T^{1/2}\Pi_{2,1,T}\left(T^{-1}Z_1'Z_j\right) + T^{1/2}\Pi_{2,2,T}\left(T^{-1}Z_2'Z_j\right) + T^{-1/2}E_2'Z_j \end{aligned}$$

for $j = 1, 2$. Using the assumptions in Scenario II, these can be rewritten as

$$\begin{aligned}\bar{B}_{1,j} &= \Pi_{1,1,T}\left(T^{-1}Z_1'Z_j\right) + T^{-1/2}C_{1,2}\left(T^{-1}Z_2'Z_j\right) + T^{-1}E_1'Z_j \\ \bar{B}_{2,j} &= C_{2,1}\left(T^{-1}Z_1'Z_j\right) + C_{2,2}\left(T^{-1}Z_2'Z_j\right) + T^{-1/2}E_2'Z_j\end{aligned}$$

for $j = 1, 2$. From the expressions above, it can be easily concluded that

$$\bar{B}_T'v \xrightarrow{d} (\bar{C}\Omega_{z,z})'v + \bar{B}_{normal}'v$$

where

$$\bar{C} = \begin{bmatrix} \Pi_{1,1} & 0 \\ C_{2,1} & C_{2,2} \end{bmatrix}.$$

By the same logic as in part (i), we can conclude that if the matrix, $\bar{C}$, is of full rank (i.e. $rank([C_{2,1}\ C_{2,2}]) = p_2$), $\bar{A}_T$ is positive definite with probability 1 and $O_p(1)$.

Substituting for $A_{i,j}$ in (36), we obtain: $|A_T| = T^{-p_2}|\bar{A}_T| = O_p(T^{-p_2})$, where the last equality follows from the properties of $\bar{A}_T$ derived above.

We now show that $\hat{\sigma}_T^2 = O_p(1)$. By definition, we have

$$T\hat{\sigma}_T^2 = u'u - 2u'X(\hat{\theta}_T - \theta_0) + (\hat{\theta}_T - \theta_0)'X'X(\hat{\theta}_T - \theta_0). \tag{38}$$

From Assumption 5 (v), it follows that $u'u = O_p(T)$, and from Assumption 5 (iii) and (v), it follows that $u'X = O_p(T)$. Therefore, we focus on $\hat{\theta}_T - \theta_0$. Let $\hat{\theta}_{T,i}$ be the 2SLS estimator of $\hat{\theta}_{0,i}$. Using the notation from the proof of part (i), $\hat{\theta}_T - \theta_0 = (B_T D_T B_T')^{-1} B_T D_T Z'u$, and so it follows that

$$\begin{bmatrix} \hat{\theta}_{T,1} - \theta_{0,1} \\ \hat{\theta}_{T,2} - \theta_{0,2} \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix} \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix} \begin{bmatrix} D_{1,1} & D_{1,2} \\ D_{2,1} & D_{2,2} \end{bmatrix} \begin{bmatrix} Z_1'u \\ Z_2'u \end{bmatrix} \tag{39}$$

where, from Dhrymes (1984)[Proposition 31], $H_{1,1} = (A_{1,1} - A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1}$, $H_{1,2} = -A_{1,1}^{-1}A_{1,2}(A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2})^{-1}$, $H_{2,1} = -A_{2,2}^{-1}A_{2,1}(A_{1,1} - A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1}$, $H_{2,2} = (A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2})^{-1}$. Using the order statements given above, it can be shown that

$H_{1,1} = O_p(1)$, $H_{1,2} = O_p(T^{1/2})$, $H_{2,1} = O_p(T^{1/2})$ and $H_{2,2} = O_p(T)$. Multiplying out (39), we obtain

$$\hat{\theta}_{T,1} - \theta_{0,1} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} H_{1,i}B_{i,j}D_{j,k}Z_k'u\,, \tag{40}$$

$$\hat{\theta}_{T,2} - \theta_{0,2} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} H_{2,i}B_{i,j}D_{j,k}Z_k'u\,. \tag{41}$$

Using the order statements given above, it follows from (40)-(41) that $\hat{\theta}_{T,1}-\theta_{0,1} = O_p(T^{-1/2})$ and $\hat{\theta}_{T,2}-\theta_{0,2} = O_p(1)$. Using these order statements along with the others above, it follows from (38) that $\hat{\sigma}_T^2 = O_p(1)$. The desired result then follows from (21).

Part (iii): It follows from Theorem 4 that $\hat{\theta}_T - \theta_0 = O_p(T^{-1/2})$. Furthermore, from Assumption 5, we have that $X'u = O_p(T)$, $X'X = O_p(T)$ and $T^{-1}u'u = \sigma_0^2 + o_p(1)$. Therefore, it follows from (38) that $\hat{\sigma}_T^2 \xrightarrow{p} \sigma_0^2$. Now consider $ln[|T^{-1}X'Z(T^{-1}Z'Z)^{-1}T^{-1}Z'X|]$. Since the $ln(.)$ is a continous function and $|T^{-1}X'Z(T^{-1}Z'Z)^{-1}T^{-1}Z'X|$ is a continuous function of the elements of $T^{-1}X'Z(T^{-1}Z'Z)^{-1}T^{-1}Z'X$, it follows from Assumption 5, Slutsky's Theorem and (32) that $ln[|T^{-1}X'Z(T^{-1}Z'Z)^{-1}T^{-1}Z'X|] \xrightarrow{p} ln[|\Omega_{1,x}'\Omega_{1,1}^{-1}\Omega_{1,x}|]$, which completes the proof.

Proof of Theorem 6

From Theorem 5(i)-(ii), it follows that $RMSC(c) \to \infty$ as $T \to \infty$ with probability 1 for all $c \in C_I \cup C_{II}$. From Theorem 5(iii), $RMSC(c) = O_p(1)$ for $c \in C_{III}$. Therefore, $lim_{T\to\infty}P(\hat{c}_T \in C_{III}) = 1$. The rest of the proof follows by the same argument as in the proof of Theorem 3.

References

Ahmed, N., and Gokhale, D. (1989). 'Entropy Expressions and Their Estimators for Multivariate Distributions', *IEEE Transactions on Information Theory*, 35: 688–692. 3

Akaike, H. (1974). 'A new look at statistical model identification', *IEEE Transactions on Automatic Control*, AC-19(6): 716–723. 3

Andrews, D. W. K. (1991). 'Heteroscedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica*, 59: 817–858. 3

——— (1999). 'Consistent moment selection procedures for Generalized Method of Moments estimation', *Econometrica*, 67: 543–564. 3, 3

——— (2000). 'Consistent moment selection procedures for GMM estimation: strong consistency and simulation results', Discussion paper, Cowles Foundation for Economics, Yale University, New Haven, CT. 3

Breusch, T., Qian, H., Schmidt, P., and Wyhowski, D. (1999). 'Redundancy of moment conditions', *Journal of Econometrics*, 91: 89–111. 2, 9, 4

Dhrymes, P. J. (1984). *Mathematics for Econometrics*. Springer Verlag, New York, NY, U. S. A., second edn. 6, 6, 6

Godambe, V. P. (1960). 'An optimum property of regular maximum likelihood estimation', *Annals of Mathematical Statistics*, 31: 1208–1211. 2, 2

Hahn, J., and Inoue, A. (2002). 'A Monte Carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators', *Econometric Reviews*, 21: 309–336. 17, 18

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press, Oxford, U.K. 2, 4, 9

Hall, A. R., and Peixe, F. P. M. (2003). 'A consistent method for the selection of relevant instruments in linear models', *Econometric Reviews*, 22: 269–288. 3, 3, 6

Hall, A. R., Rudebusch, G., and Wilcox, D. (1996). 'Judging instrument relevance in instrumental variables estimation', *International Economic Review*, 37: 283–298. 19

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. 3, 10

Hansen, L. P. (1982). 'Large sample properties of Generalized Method of Moments estimators', *Econometrica*, 50: 1029–1054. 2

Jana, K. (2005). 'Canonical correlations and instrument selection in econometrics', Ph.D. thesis, Department of Economics, North Carolina State University, Raleigh, NC, USA. 2, 6

Kitamura, Y. (1997). 'Empirical likelihood methods with weakly dependent processes', *Annals of Statistics*, 25: 2084–2102. 2

Maasoumi, E. (1993). 'A compendium to information theory in economics and econometrics', *Econometric Reviews*, 12: 137–181. 5

Magnus, J. R., and Neudecker, H. (1991). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York, NY. 6

Nelson, C. R., and Startz, R. (1990). 'The distribution of the instrumental varaibles estimator and its t ratio when the instrument is a poor one', *Journal of Business*, 63: S125–S140. 19

Neyman, J. (1949). 'Contributions to the theory of the chi-squared test', in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 239–273. University of California Press, Berkeley, CA, USA. 2

Owen, A. B. (1988). 'Empirical likelihood ratio confidence intervals for a single functional', *Biometrika*, 75: 237–249. 2

——— (2001). *Empirical Likelihood*. Chapman Hall, London, U.K. 2

Qin, J., and Lawless, J. (1994). 'Empirical likelihood and generalized estimating equations', *Annals of Statistics*, 22: 300–325. 2

Qu, A., Lindsay, B., and Li, B. (2000). 'Improving generalized estimating inference equations using quadratic inference functions', *Biometrika*, 87: 823–836. 2

Rao, C. R. (1973). *Linear statistical inference and its applications*. Wiley, New York, NY USA, second edn. 6

Schwarz, G. (1978). 'Estimating the dimension of a model', *Annals of Statistics*, 6: 461–464. 3

Staiger, D., and Stock, J. H. (1997). 'Instrumental variables regression with weak instruments', *Econometrica*, 65: 557–586. 4, 6

Tauchen, G. (1985). 'Diagnostic testing and evaluation of maximum likelihood models', *Journal of Econometrics*, 30: 415–443. 2

Zellner, A. (2003). 'Some aspects of the hsitory of Bayesian information processing', Discussion paper, Graduate School of Business, University of Chicago, Chicago IL. 1

Zivot, E., Startz, R., and Nelson, C. R. (2003). 'Inference in partially identified instrumental variables regression with weak instruments', Discussion paper, Department of Economics, University of Washington, Seattle WA. 14

Table 1: Median Bias of 2SLS Estimator

T	R_f^2	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$	$q=7$	$q=8$	$q=9$	$q=10$	$q=11$	$q=12$	$q=\hat{q}$
100	0.5	0.001	0.007	0.011	0.016	0.021	0.026	0.030	0.035	0.039	0.042	0.047	0.051	0.001
100	0.1	0.004	0.047	0.084	0.115	0.143	0.163	0.184	0.203	0.218	0.232	0.245	0.256	0.097
500	0.5	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010	0.011	0.000
500	0.1	0.001	0.010	0.018	0.026	0.034	0.043	0.050	0.057	0.064	0.070	0.077	0.084	0.004

Table 2: Coverage Probabilities of 90 Percent Confidence Intervals for 2SLS Estimator

T	R_f^2	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$	$q=7$	$q=8$	$q=9$	$q=10$	$q=11$	$q=12$	$q=\hat{q}$
100	0.5	0.901	0.897	0.896	0.891	0.888	0.881	0.873	0.866	0.858	0.850	0.841	0.832	0.901
100	0.1	0.920	0.899	0.870	0.845	0.814	0.785	0.752	0.721	0.691	0.659	0.626	0.594	0.881
500	0.5	0.900	0.900	0.898	0.896	0.896	0.895	0.894	0.891	0.890	0.889	0.888	0.885	0.900
500	0.1	0.907	0.901	0.895	0.890	0.880	0.872	0.861	0.852	0.838	0.824	0.812	0.796	0.908

Table 3: Summary Statistics of $\hat{q}$

		2SLS			
T	R_f^2	mean	median	mode	variance
100	0.5	1.000	1.000	1.000	0.000
100	0.1	1.785	1.000	1.000	2.293
500	0.5	1.000	1.000	1.000	0.000
500	0.1	1.052	1.000	1.000	0.090

Table 4: Properties of *RMSC(c)*

$R_f^2 = 0.1$	*T=100*			*T=500*		
Empirical Selection Probabilities:						
inst.$\diagdown\sigma_{ue}$	*0.1*	*0.5*	*0.9*	*0.1*	*0.5*	*0.9*
1R	0.220	0.196	0.103	0.109	0.152	0.151
2R	0.291	0.220	0.072	0.836	0.690	0.420
1R/I	0.329	0.416	0.571	0.012	0.046	0.216
*2R/I**	0.138	0.136	0.059	0.043	0.113	0.201
I	0.021	0.032	0.194	0.000	0.000	0.012
All	0.000	0.000	0.000	0.000	0.000	0.000
Sampling Properties of Post-Selection Estimator:						
Med Bias	0.042	0.210	0.477	0.005	0.033	0.084
Cov Rate	0.947	0.804	0.253	0.919	0.897	0.766

$R_f^2 = 0.5$	*T=100*			*T=500*		
Empirical Selection Probabilities:						
inst.$\diagdown\sigma_{ue}$	*0.1*	*0.5*	*0.9*	*0.1*	*0.5*	*0.9*
1R	0.290	0.342	0.364	0.001	0.005	0.033
2R	0.709	0.654	0.593	0.999	0.995	0.967
1R/I	0.000	0.002	0.032	0.000	0.000	0.000
*2R/I**	0.000	0.001	0.011	0.000	0.000	0.000
I	0.000	0.000	0.000	0.000	0.000	0.000
All	0.000	0.000	0.000	0.000	0.000	0.000
Sampling Properties of Post-Selection Estimator:						
Med Bias	0.003	0.017	0.027	0.001	0.001	0.002
Cov Rate	0.906	0.903	0.876	0.903	0.908	0.898

Notes: *inst.* stands for instrument combination; $1R$ denotes the cases in which $c = (a, 0_6')'$ for $a \in \{(1,0),(0,1)\}$, $2R$ denotes the case in which $c = (1, 1, 0_6')'$, $1R/I$ denotes the cases in which $c = (a', b')'$ for a given above and $b \neq 0_6$, $2R/I^*$ denotes the cases in which $c = (1, 1, d')'$ and $d \neq 0_6$ *or* ι_6, I denotes the cases in which $c = (0, 0, b')'$ for b given above, and *all* denotes the case in which $c = \iota_8'$.