Contemporaneous and long run canonical correlations in the linear IV model: Implications for instrument selection

Gunce Eryuruk a, Alastair R. Hall b,⁎, Kalidas Jana c

a ITAM, Mexico
b Economics, School of Social Sciences, University of Manchester, Manchester M13 9PL, UK
c University of Texas at Brownsville, United States

Abstract

In the normal linear simultaneous equations model, we demonstrate a close relationship between two recently proposed methods of instrument selection by presenting a fundamental relationship between the two sets of canonical correlations upon which the methods are based.

Keywords:
Contemporaneous canonical correlations
Long run canonical correlations
Instrument selection
Two-stage least squares

JEL classification:
C13
C15
C30

1. Introduction

Ordinary Least Squares (OLS) estimation yields inconsistent estimators of the parameters of linear regression models when regressors are correlated with errors of the model. In such models a popular method for obtaining consistent estimators is application of the Instrumental Variables (IV) method. To implement the IV method in practice, the researcher must choose a set of instruments. In the past, such choices have been informal at best. Recently, a number of formal criteria have been proposed in the literature to remedy that problem. This paper relates to two among these proposed criteria, namely, the Canonical Correlations Information Criterion (CCIC) of (Hall and Peixe, 2003) and the Relevant Moments Selection Criterion (RMSC) of Hall et al. (2007). The objective of Hall and Peixe (2003) and Hall et al. (2007) is to achieve an improved quality of asymptotic approximation to the finite sample behavior of the estimators. They gain this objective by eliminating the redundant moment conditions based on certain canonical correlations: CCIC exploits explicitly the canonical correlations (CCs) between the regressors and instruments; RMSC exploits implicitly the long run canonical correlations (LRCCs) between the unknown true score vector and the product of the instrument vector and error.

In this paper, we establish an interesting relation between LRCCs and CCs in a linear simultaneous equations model that helps explain the connection between CCIC and RMSC in this model. We further use the aforementioned result to reveal an interesting structure to the information measure that underlies RMSC, and also to relate RMSC to CCIC.

2. Canonical correlations and information in IV estimation

It is noted in the Introduction that, in the linear model, RMSC implicitly exploits the information in the LRCCs between the score and product of the instrument and error. In this section, we derive an explicit representation for these LRCCs in the linear simultaneous equations model with normal errors. This representation turns out to involve the CCs between the regressors and instruments, and we explore its implications for the information metric upon which RMSC is based.

To begin, it is useful to formally define CCs and LRCCs.1

Definition 1. Canonical correlations

1 CCs are introduced by Hotelling (1935); LRCCs are introduced (to our knowledge) in Hall et al. (2007). Note that, for ease of presentation, it is taken for granted in Definitions 1 and 2 that all expectations and inverses exist.
Let $v_{ti} = \mathbf{k}_i \times 1$ vectors for $i = 1, 2$ and $m = \min(k_i k_2)$. Let $v_i = (v_{ti}, v_{ti}^*)$, $[v_i; t = 1, 2, \ldots, T]$, be a covariance stationary process with $\Sigma_i = \text{Var}[v_i]$ where

$$
\Sigma_i = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{bmatrix}
$$

in the obvious notation. The population canonical correlations between $v_{t1}$ and $v_{t2}$ are denoted by $\{r_{ij}; i = 1, 2, \ldots, m\}$, where by convention $r_{ji} \geq 0$ for $i = 1, 2, \ldots, m$ and $r_{ji} \geq r_{ji-1}$ for $i = 1, 2, \ldots, m-1$, and have the following properties: (i) $r_{ij}^2$ are the m largest solutions to the determinantal equation $|\Sigma_{11} - r_{ij}^2 \Sigma_{12} - r_{ij}^2 \Sigma_{21} - r_{ij}^2 \Sigma_{22}| = 0$; (ii) $r_{ij} = a_i \sum j = b_i$, where $a_i$ and $b_i$ satisfy $\sum a_i^2 = \sum b_i^2 = \sum r_{ij}^2 = \sum r_{ij}^2 = 0$ and $(\sum a_i^2, \sum a_i, \sum b_i, \sum b_i^2) = 0$ for $i = 1, 2, \ldots, m$.

**Definition 2.** Long run canonical correlations

Let $v_i$ be as in **Definition 1** and define $\Omega_i = \lim_{T \to \infty} \text{Var}[(T^{-1/2} \sum_{t=1}^T v_i)]$ where

$$
\Omega_i = \begin{bmatrix}
\Omega_{11} & \Omega_{12} \\
\Omega_{21} & \Omega_{22}
\end{bmatrix}
$$

in the obvious notation. The long run population canonical correlations between $v_{t1}$ and $v_{t2}$ are denoted by $\{r_{ij}; i = 1, 2, \ldots, m\}$ and satisfy the equations given in **Definition 1** (i) and (ii) when $\Sigma_i$ is replaced by $\Omega_i$ for all $i = 1, 2$.

A comparison of Definitions 1 and 2 indicates that the LRCCs between $v_{t1}$ and $v_{t2}$ are the limiting CCs between $T^{-1/2} \sum_{t=1}^T v_{t1}$ and $T^{-1/2} \sum_{t=1}^T v_{t2}$.

For the rest of this paper, we consider the following linear simultaneous equations model

$$
y_{it} = x_{it}' \theta_0 + u_{it}, \quad (i = 1, 2) \quad (1)$$

$$x_{it} = \pi_{it} z_{it} + e_{it}, \quad t = 1, 2, \ldots, T \quad (p \times 1)$$

where

$$
\begin{bmatrix}
u_i \\
e_i
\end{bmatrix} | z_i \sim \mathcal{N}_p \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma & \Psi \\ \Psi & \Sigma \end{bmatrix} \right) \quad (2)
$$

with

$$
\sum_{(p+1) \times (p+1)} = \begin{bmatrix}
\alpha_i^2 & \sum_{i=1}^p \alpha_i \\
\sum_{i=1}^p \alpha_i & \sum_{i=1}^p \alpha_i
\end{bmatrix} \quad (3)
$$

It is mentioned above that the LRCCs of interest involve the score function. Following (Hall et al., 2007) the relevant score is the one associated with the model in Eqs. (1)–(3) for the case in which the only unknown parameters contained in $\theta_0$. Accordingly, we define $w_i = (y_{i1}, x_{i1})$ and $p(w_i | z_i, \theta)$ to be the conditional probability density function (pdf) of $w_i$ given $z_i$ implied by Eqs. (1)–(3) assuming $\Sigma$ and $\pi_{it}$ are known, and $k_i$ is the associated score with respect to $\theta$ evaluated at $\theta = \theta_0$, that is $s_i(\theta_0) = \left\{ \partial \ln p(w_i | z_i, \theta) / \partial \theta \right\} |_{\theta = \theta_0}$. Let $\{p^2; ; 1, 2, \ldots, p\}$ denote the population squared LRCCs between the score vector $S(\theta_0)$ and the vector $z_{it}(\theta_0)$.

In addition to the conditions above, we impose the following regularity conditions.

**Assumption 1.** (i) $\theta_0$ is an interior point of the parameter space $\Theta$; (ii) $z_i$ is covariance stationary, weakly exogenous for estimation of $\theta_0$; and $E[z_i z_i^*]$ exists, is finite and is positive definite; (iii) $\pi_{it} \pi_{it}^* < \infty$; (iv) $\Sigma$ is finite and positive definite.

The following theorem specifies the form of the aforementioned LRCCs in this model.

**Theorem 1.** Let the data be generated by Eqs. (1)–(3) and Assumption 1 hold. Then the population squared LRCCs between the score vector $S(\theta_0)$ and the vector $z_{it}(\theta_0)$ are given by

$$
\rho_i^2 = \rho_i^2 r_i^2 (z), \quad i = 1, 2, \ldots, p. \quad (4)
$$

where $R_i^2$ is the population squared multiple correlation coefficient between the structural equation error $u_i$ and the reduced form equation error $e_i$, and $r_i^2(z)$ is the ith population squared CC between $x_i$ and $z_i$.

The proof is available upon request. We now explore some interesting implications of **Theorem 1**.

**Remark 1.** Since $0 \leq (1 - R_i^2) \leq 1$, it follows that $\rho_i^2 \leq r_i^2(z)$. However, in other words, the population squared LRCCs between the score vector $S(\theta_0)$ and the vector $z_{it}(\theta_0)$ are bounded by the corresponding population squared canonical correlation between $x_i$ and $z_i$. Furthermore, if $\theta_0$ is endogenous, the CCs between $x_i$ and $z_i$ are positively related to the multiple correlation between $u_i$ and $e_i$.

**Remark 2.** If $p = 1$ and $x_i$ is a scalar—then $R_i^2$ is the squared correlation between the structural equation error, $u_i$, and the scalar reduced form error, $e_i$, and $1 - R_i^2$ can be interpreted as a measure of endogeneity of $x_i$. For, if $1 - R_i^2 = 0$, then $u_i$ and $e_i$ are uncorrelated and hence $x_i$ is exogenous; if $1 - R_i^2 < 1$, then $u_i$ and $e_i$ are correlated and hence $x_i$ is endogenous. In this case, it can be shown that, ceteris paribus, the correlation between $x_i$ and $u_i$ is positively related to $R_i^2$. Therefore, the magnitudes of the LRCCs are inversely related to the degree of endogeneity of the regressor, ceteris paribus.

**Remark 3.** Hall et al. (2007) argue that the entropy of the limiting distribution of the GMM—or in this case IV—estimator provides a natural metric for the information about the parameters in moment based estimation. They further show that this entropy only depends on the choice of moment condition via the LRCCs between the score and the function of the data and parameter vector appearing in the moment condition. Combining Hall et al. (2007) [Eq. (6)] with Eq. (4), it follows that, in the model considered here, the entropy of the limiting distribution of the IV estimator of $\theta_0$, $ent(z)$ say, has the following structure

$$
ent(z) = 0.5 \rho_i^2 [1 + \ln(2\pi)] + 0.5 \ln[|S(z)|] - 0.5 \rho_i^2(1 - R_i^2) \quad (5)
$$

Note that the first two terms on the right hand side of Eq. (5) represent the entropy of the MLE based on the true pdf $p(w_i | z_i, \theta_0)$, that is the GMM estimator based on the true score. This MLE is, of course, infeasible because in general $\Sigma$ and $\pi_{it}$ are unknown; however, this infeasible MLE can be regarded as optimal within the class of GMM estimators of $\theta_0$. Since squared correlations lie in the unit interval, the third and fourth terms are positive and so represent the efficiency loss relative to this aforementioned infeasible optimal estimator. Interestingly, Eq. (5) reveals that this efficiency loss depends positively on the multiple correlation between $u_i$ and $e_i$ and negatively on the CCs between $x_i$ and $z_i$.

**Remark 4.** The results in **Theorem 1** point to a relationship between two recently proposed methods of instruments selection, the CCIC of (Hall and Peixe, 2003) and the RMSC of Hall et al. (2007). Both criteria are designed to eliminate redundant moment conditions based on certain canonical correlations: CCIC exploits explicitly the canonical
correlations (CCs) between the regressors and instruments; RMSC exploits implicitly the long run canonical correlations (LRCCs) between the unknown true score vector and the product of the instrument vector and error. The results in Theorem 1 indicate that, for the model under consideration here, the two criteria are fundamentally linked as they are both driven by the canonical correlations between the regressors and instruments. Using $p = 1$, that is, only a single endogenous regressor in models (1)–(3) considered above, this fundamental link can be easily seen as follows.

CCIC of (Hall and Peixé, 2003) is defined to be

$$\text{CCIC}(c) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|c|} \ln |1 - r^2_{z(c)}| \quad (6)$$

where the statistic

$$\Xi_T(c) = T \sum_{t=1}^{p} \ln (1 - r^2_{z(c)}) \quad (7)$$
captures the sample information; the $q \times 1$ selection vector $c$ denotes, in the notation of (Andrews, 1999), which elements of the instrument vector $z$ are included in a particular moment condition: if $c_j = 1$ then the $j$th element of $z$ is included, if $c_j = 0$ then the $j$th element of $z$ is excluded; $|c| = c'c$ equals the number of elements in the instrument vector $z(c)$ and $P(T|c)$ is a “penalty” term.

If the regressor $x_t$ is a scalar, CCIC involves only one sample squared canonical correlation, $r^2_{z(c)}$, which is equal to the squared multiple correlation coefficient, also commonly known as the coefficient of determination. Specializing the definitions in Eq. (6) to the case of a single endogenous regressor and the penalty term associated with Bayesian Information Criterion, CCIC takes the form

$$\text{CCIC}(c) = T \ln |1 - r^2_{z(c)}| + (\langle |c| \rangle - 1) \ln T. \quad (8)$$

RMSC of (Hall et al., 2007) is given by

$$\text{RMSC}(c) = \ln |\hat{V}_{cT}(c)| + \kappa(\langle |c| \rangle, T) \quad (9)$$

where $\hat{V}_{cT}(c)$ denotes a consistent estimator of the asymptotic variance $V(c)$ of the GMM estimator $\hat{b}_T(c)$ and $\kappa(\langle |c| \rangle, T)$ is a deterministic penalty function. Specializing Eq. (9) to our simple linear IV model yields RMSC criterion

$$\text{RMSC}(c) = \ln \left( |\hat{b}_T(c)|^2 \sum_{t=1}^{T} x_t z_t(c) \sum_{t=1}^{T} z_t(c) z_t(c)' \right)^{-1} \sum_{t=1}^{T} z_t(c) x_t(c)^{-1} \left| z_t(c) x_t(c)^{-1} \right|^{-1} \right)$$

$$+ \frac{(\langle |c| \rangle - 1)}{\sqrt{T}} \ln \sqrt{T}$$

Using similar arguments to those behind Theorem 1, it can be shown that

$$\text{RMSC}(c) = b_T \ln |r^2_{z(c)}| + \frac{(\langle |c| \rangle - 1)}{\sqrt{T}} \ln \sqrt{T} \quad (10)$$

where $b_T$ is $O_p(1)$ and independent of $c$. It can be seen from Eq. (10) that instrument selection based on RMSC depends purely on the data via the squared canonical correlation which in this case is the squared multiple correlation coefficient, $r^2_{z(c)}$. Thus both CCIC and RMSC base instrument selection on $r^2_{z(c)}$ but this correlation is compared to a different deterministic function in each case. For example, suppose we consider the choice between two instrument vectors: $z_1(c)$ and $z_2(c)$ where the latter includes all the instruments in the former plus one more. The vector $z_2(c)$ is selected over $z_1(c)$ according to CCIC if

$$1 - \frac{r^2_{z_1(c)}}{r^2_{z_2(c)}} < \exp \left[ - \left( \frac{\langle |c| \rangle - 1}{\sqrt{T}} \right) \right]. \quad (11)$$

and $z_2(c)$ is selected over $z_1(c)$ according to RMSC if

$$\frac{r^2_{z_1(c)}}{r^2_{z_2(c)}} > \exp \left[ \frac{0.5(\langle |c| \rangle - 1)}{\sqrt{T}} \right]. \quad (12)$$

Calculations based on Eqs. (11) and (12) reveal the following about the relative properties of the two criteria in this setting for $30 < T < 1000$. If $r^2_{z_2(c)} = (0.01, 0.05)$ then as $\eta = r^2_{z_2(c)} - r^2_{z_1(c)}$ increases from zero we can divide the range of $\eta$ into three intervals: small values $(\eta \in [0, \eta_1])$ for which neither CCIC nor RMSC indicate the additional instrument should be included; then a range of values $(\eta \in [\eta_1, \eta_2])$ for which RMSC indicates that the additional instrument should be included but CCIC does not; and finally a range of values $(\eta > \eta_2)$ for which both criteria indicate the additional instrument should be included. If $r^2_{z_2} = 0.1$ then the range of $\eta$ can again be divided into three intervals but this time the qualitative decision in the middle range depends on $T$: for $T > 266$ the middle range involves values of $\eta$ for which RMSC indicates inclusion of the additional instrument but CCIC does not; for $T > 266$ this is reversed. For $r^2_{z_2} \in (0.25, 0.5)$ and $T > 30$, the middle range involves values of $\eta$ for which CCIC indicates inclusion but RMSC does not. These values of $\eta_1$ depend on both $T$ and $r^2_{z_2(c)}$. We note that for $r^2_{z_2(c)} = (0.01, 0.05)$, there are values of $T$ greater than 1000 at which the decision in the middle range is reversed. However, for these samples sizes, the width of the middle range is very small. The latter reflects the consistency of both methods which implies that the first two intervals combined, $(0, \eta_2]$, become empty so that both methods indicate the additional instrument should be included for any $\eta > 0$ and any $r^2_{z_2} \in [0, 1]$ in the limit with probability one.

References


