

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/econlet

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

ARTICLE INFO

Article history:

Received 18 July 2008

Received in revised form 18 May 2009

Accepted 10 June 2009

Available online 21 June 2009

Keywords:

Contemporaneous canonical correlations

Long run canonical correlations

Instrument selection

Two-stage least squares

JEL classification:

C13

C15

C30

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.

© 2009 Elsevier B.V. All rights reserved.

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.

* Corresponding author. Tel.: +44 1612754875; fax: +44 1612754812.
E-mail address: alastair.hall@manchester.ac.uk (A.R. Hall).

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 Peixe, 2003) is defined to be

$$CCIC(c) = \Xi_T(c) + P(T, |c|) \tag{6}$$

where the statistic

$$\Xi_T(c) = T \sum_{i=1}^p \ln[1 - r_{iT}^2(c)] \tag{7}$$

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_t are included in a particular moment condition: if $c_j = 1$ then the j th element of z_t is included, if $c_j = 0$ then the j th element of z_t is excluded; $|c| = c'c$ equals the number of elements in the instrument vector $z_t(c)$ 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_T^2 , 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

$$CCIC(c) = T \ln[1 - r_T^2(c)] + (|c| - 1) \ln T. \tag{8}$$

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

$$RMSC(c) = \ln[|\hat{V}_{\theta, T}(c)|] + \kappa(|c|, T) \tag{9}$$

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

$$RMSC(c) = \ln \left(\left| \hat{\sigma}_u^2 \left[\sum_{t=1}^T x_t z_t(c)' \left\{ \sum_{t=1}^T z_t(c) z_t(c)' \right\}^{-1} \sum_{t=1}^T z_t(c) x_t(c) \right]^{-1} \right| \right) + \frac{(|c| - 1)}{\sqrt{T}} \ln \sqrt{T}$$

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

$$RMSC(c) = b_T - \ln[r_T^2(c)] + \frac{(|c| - 1)}{\sqrt{T}} \ln \sqrt{T} \tag{10}$$

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, $r_T^2(c)$. Thus both CCIC and RMSC base instrument selection on $r_T^2(c)$ but this correlation is compared to a different deterministic function in each case. For example, suppose we consider the choice between two instrument vectors: $z_t(c_1)$ and $z_t(c_2)$ where the latter includes all the instruments in the former plus one more. The vector $z_t(c_2)$ is selected over $z_t(c_1)$ according to CCIC if

$$\frac{1 - r_T^2(c_2)}{1 - r_T^2(c_1)} < \exp\{-[\ln T] / T\} \tag{11}$$

and $z_t(c_2)$ is selected over $z_t(c_1)$ according to RMSC if

$$\frac{r_T^2(c_2)}{r_T^2(c_1)} > \exp\{0.5[\ln T]\} / \sqrt{T}. \tag{12}$$

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 $r_T^2(c_1) \in \{0.01, 0.05\}$ then as $\eta = r_T^2(c_2) - r_T^2(c_1)$ increases from zero we can divide the range of η in three intervals: small values ($\eta \in [0, n_1)$) for which neither CCIC nor RMSC indicate the additional instrument should be included; then a range of values ($\eta \in [n_1, n_2)$) for which RMSC indicates that the additional instrument should be included but CCIC does not; and finally a range of values ($\eta > n_2$) for which both criteria indicate the additional instrument should be included. If $r_T^2 = 0.1$ then the range of η 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 η for which RMSC indicates inclusion of the additional instrument but CCIC does not, but for $T \geq 266$ this is reversed. For $r_T^2 \in \{0.25, 0.5\}$ and $T > 30$, the middle range involves values of η for which CCIC indicates inclusion but RMSC does not. These values of $\{n_i\}$ depend on both T and $r_T^2(c_1)$. We note that for $r_T^2(c_1) \in \{0.01, 0.05\}$, 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, $(0, n_2]$, become empty so that both methods indicate the additional instrument should be included for any $\eta > 0$ and any $r_T^2(c_1) < 1 - \eta$ in the limit with probability one.

References

- Andrews, D.W.K., 1999. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* 543–564.
- Hall, A.R., Inoue, A., Jana, K., Shin, C., 2007. Information in generalized method of moments estimation and entropy based moment selection. *Journal of Econometrics* 488–512.
- Hall, A.R., Peixe, F.P.M., 2003. A consistent method for the selection of relevant instruments. *Econometric Reviews* 3, 269–287.
- Hotelling, H., 1935. The most predictable criterion. *Journal of Educational Psychology* 139–142.