Cross-Entropy Estimation of Linear Cointegrated Equations

Balcombe, Kelvin

University of Reading

29 January 2006

Cross-Entropy Estimation of Linear Cointegrated Equations

Kelvin Balcombe

Dept of Agricultural and Food Economics,
University of Reading.
[1]

ABSTRACT

The cross-entropy approach is extended to the estimation of cointegrated equations. The entropy estimators for an appropriately constructed moment form, are asymptotically equivalent to Fully Modified estimators since they converge to these estimates sufficiently quickly. The performance of the entropy estimators are examined by using some Monte Carlo trials, and in an applied example for the estimation of a production function for South African agriculture.

Key Words: FM-SUR, Entropy

JEL Classification: C32

[1]Email: K.Balcombe@imperial.ac.uk

1. Introduction

The publication of the book by Golan Judge and Miller (1996) (henceforth GJM), has promoted a renewed surge of interest (see AJAE 1999, Vol 3, and Journal of Econometrics 2002, Vol 107) in 'information based' estimation using the entropy measure of Shannon (1948) and Kullback-Liebler Information measure (1959). As Golan (2002) outlines, the objective of the entropy approach is to make use of partial or incomplete information. Entropy can be used in order to make minimal assumptions about the data generating process, or it can be used to integrate prior and sample information.

The aim of this paper is to develop a 'double support' approach to cross-entropy for the estimation of a cointegrated system. Since the Fully-Modified (FM) (Phillips and Hansen, 1990) estimator can be expressed as a linear solution for a moment form, given estimates of the long-run covariance matrices, entropy can be used to estimate the parameters of this system. Given prior knowledge, entropy has the potential to yield improved estimates in finite samples, with tests that have better empirical size compared to using either a simple entropy approach or a Fully Modified approach alone. In order to demonstrate the utility of the approach a small Monte-Carlo study is undertaken, along with an applied application to the estimation of a Cobb-Douglas production function within South African Agriculture.

2. Overview

There is now a considerable body of work which gives a philosophical foundation to the use of entropy as an 'extremum' criteria (Zellner, 1996). GJM outline how entropy can be used to estimate parameters in several ways. These include direct and dual approaches. The entropy approach can be applied to a wide class of models including Seemingly Unrelated Regressions (SUR, see also Harmon et al., 1998), Simultaneous Systems (Marsh et al., 1998) and systems with censored or multinomial data (Golan, et al., 1997, 1999).

The asymptotic normality and consistency of entropy estimators has been proved under the assumption that (inter-alia) the first moment matrix of the explanatory variables converges to a constant positive definite matrix. However, to the author's knowledge, there has been no work which has examined entropy estimation in the context of linear cointegrated systems. Therefore, this paper examines entropy estimation within SURs where the explanatory variables are I(1). The estimation of SURs using entropy under the assumption of stationary regressors has already been dealt with (GJM, chapter 11. and Harmon, et al., 1998). Here, the entropy approach is extended to cointegrated systems in a 'two-boundary' setting using the 'moment form'. This paper restricts its attention to the case where upper and lower boundaries are set for each of the parameters along with a prior expected value within this support. In addition, all equation

errors are assumed to be within a symmetric interval around zero, with the prior expectation for each of the errors being zero. The GJM framework allows for multiple supports for each parameter/error, and also allows non-zero prior expectations to be set for the errors. While the approach employed here is a considerable simplification of the formulations in GJM, it offers a tractable solution with entropy expressed in terms of the data, parameters, and expected values. It also enables a direct comparison of the relationship between entropy estimators and conventional Seemingly Unrelated Regression (SUR), and FM (and FM-SUR) estimators (Phillips and Hansen, 1990; Moon, 1999, and Balcombe and Tiffin, 2001).

There are insights to be gained from exploring the relationship between entropy based estimators with other estimators (Prekel, 1998). The computation of the entropy based estimators can be improved by utilising these relationships. Moreover, the asymptotic distribution of an estimator may be deduced from its convergence (and rate of convergence) to other estimators.

This paper notes some results for a 'pure inverse' problem in Section 3. It examines the relationship between the simple inverse solution for the parameters in the pure inverse problem, and that of entropy estimates. It outlines the conditions under which the matrices and vectors in the pure inverse form yield entropy estimates that converge to the inverse solution given finite supports for the parameters and equations errors. These conditions are subsequently related to the SUR and FM estimators.

Alternative entropy formulations may not require finite supports (Golan and Gzyl, 1999). However, providing the supports are made wide enough, entropy estimates will almost certainly exist and converge in distribution and this assumption does not present significant problems on a practical level. Under certain conditions entropy forms of FM and SUR are asymptotically equivalent. However, the properties of the entropy estimates may be superior in finite samples. This is particularly important with regard to the FM estimates which have excellent asymptotic properties, but do not always display these qualities in small samples.

3. The Pure Inverse Problem

Notation

In the following, the Greek letter β will be reserved for the population value of a $1 \times K$ parameter vector in a linear model. The letter b will denote an arbitrary vector in $\mathbb{R}^K$. $\hat{b}$ will always refer to the 'OLS estimate' of the parameter vector β, and $\tilde{b}$ will refer to the cross-entropy estimator of β which will be defined in the subsequent sections. $v(b)$ will also be used to denote the estimate of the residual in a model for some value of b replacing β. The notation $b > \beta$, defines every element of b to be larger than the associated element of β. For conciseness, we will use $\hat{v} = v\left(\hat{b}\right)$, $\tilde{v} = v\left(\tilde{b}\right)$

and $v = v(\beta)$ .The letter $\mathbf{b}$ (in bold) will always refer to a vector which plays the role of a bound or 'support'. The notation $\xrightarrow{d}$ and $\xrightarrow{p}$ will be used to denote convergence in distribution and probability respectively and $\mathcal{M}^{\mathcal{N}}(\mu, \Omega)$ will be used to denote the multivariate normal distribution with mean μ and covariance matrix Ω.

Examine the linear system

$$\underset{K\times 1}{\gamma} = \underset{K\times K}{M}\ \underset{K\times 1}{\beta} + \underset{K\times 1}{v} \tag{3.1}$$

where M and γ are observable matrices with M being invertible (a restrictive assumption that is made throughout the paper). Assume that $\underset{K\times 1}{v}$ is treated as a random error with a mean of zero and an identity covariance matrix.

The OLS estimate of β is simply $\hat{b} = M^{-1}\gamma$. Cross-entropy is not directly defined in terms of parameters and errors. Rather, it is defined on a set of probabilities $p_1, ...p_k$ where $\sum p_i = 1$. If $p_1^*, ...p_k^*$ represent a set of prior expectations of the probabilities, then cross-entropy is defined as $C = \sum_{i=1}^{k} p_i \ln\left(\frac{p_i}{p_i^*}\right)$. Cross-entropy is a measure of the divergence of the probabilities from the prior p_i^*. The essential idea behind using cross-entropy estimation is to find estimates of the probabilities that minimise the divergence of the probabilities from their priors, as measured by cross-entropy, subject to a set of constraints that arise from a model and data.

The objective cross-entropy function is sum of two cross-entropy functions, one for the errors (measuring their divergence from zero), and one for the parameters (b) (measuring their divergence from prior or 'expected' values). If the prior values for b are completely compatible with the data, then both cross-entropy functions would be at their minimal values. However, for a given data set (i.e. values of γ and M in [3.1]), the parameters (b) consistent with zero errors are likely to differ from their priors. Likewise, if the parameters were set at their prior values, the errors (v) would not be zero.

A 'two boundary' cross-entropy formulation is outlined more formally below. Let $\mathbf{b}_l$ and $\mathbf{b}_u$ be $(K \times 1)$ vectors, where $\mathbf{b}_u > \mathbf{b}_l$ (this notation denoting that every element of $\mathbf{b}_u$ is larger than the associated element of $\mathbf{b}_l$). It is assumed that $\beta \in (\mathbf{b}_l, \mathbf{b}_u)$ where $(\mathbf{b}_l, \mathbf{b}_u)$ denotes the set

$$\left\{(b_1....,b_K)' : b_k = p_k\mathbf{b}_{l,k} + (1-p_k)\mathbf{b}_{u,k}, p_k \in (0,1)\right\} \tag{3.2}$$

Put simply, the vector β is specified so as to lie within a predetermined interval. Likewise, it is assumed that the errors are symmetrically distributed around zero and supported by the set $(-\mathbf{s}, \mathbf{s})$ ($\mathbf{s}$ being a $K \times 1$vector $\mathbf{s}'=(s, ...s)$). The elements of β and v can therefore be expressed as

$$b_k = p_k\mathbf{b}_{u,k} + (1-p_k)\mathbf{b}_{l,k} \tag{3.3}$$

$$v_k = w_k s - (1 - w_k)\, s \tag{3.4}$$

where $p' = (p_1, ...p_K)$ and $w' = (w_1, ...w_K)$ and all elements of p and w are between zero and one. There may be prior knowledge about the parameter values in the form of an 'expected value' for each of the estimates (b_k^*) which are within the supports. The associated values p_k^* are the probabilities that solve the equation (given $\mathbf{b}_{u,k}$, $\mathbf{b}_{l,k}$ and b_k^*)

$$p_k^* = \frac{b_k^* - \mathbf{b}_{l,k}}{\mathbf{b}_{u,k} - \mathbf{b}_{l,k}} \tag{3.5}$$

If b and v are defined in terms of p and w as in [3.3] and [3.4], then equations may be reversed so as to express the p and w in terms of b and v as below

$$p_k = \frac{b_k - \mathbf{b}_{l,k}}{\mathbf{b}_{u,k} - \mathbf{b}_{l,k}} \tag{3.6}$$

and

$$\begin{aligned} w_k &= \left(\frac{v_k + s}{2s}\right) \\ &\quad where \\ v &= \gamma - Mb. \end{aligned} \tag{3.7}$$

Throughout this article, we will refer to the problem in terms of one to be maximised (maximising negative cross-entropy). The objective function [3.8] (negative cross-entropy) can be expressed in terms of p and w as (treating $s, \mathbf{b}_l, \mathbf{b}_u, M, \gamma, p^*$ as constants which underly the construction of $f^*(.)$ and $g^*(.)$) :

$$E^*(p, w) = f^*(w) + g^*(p) \tag{3.8}$$

where

$$\begin{aligned} g^*(p) &= -\sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{p_k^*}\right) - \sum_{k=1}^{K} (1 - p_k) \ln\left(\frac{1 - p_k}{1 - p_k^*}\right) \\ f^*(w) &= -\sum_{k=1}^{K} w_k \ln(w_k) - \sum_{k=1}^{K} (1 - w_k) \ln(1 - w_k) \end{aligned} \tag{3.9}$$

The second of these functions is equal to $-ln(\frac{1}{2})K$ minus cross-entropy if the priors are $w_k^* = \frac{1}{2}$ for all K. However, since p_k, p_k^* and w_k can be expressed as in [3.5],[3.6] and [3.7], we can substitute in these quantities so as to obtain

$$\begin{aligned} E(b) &= f^*(p(b)) + g^*(w(b)) \\ &= f(b) + g(b) \end{aligned} \tag{3.10}$$

The cross-entropy estimates for b and v ($\tilde{b}$ and $\tilde{v}$) are then obtained by maximising [3.8] and equivalently [3.10] subject to the constraints in [3.1]. An insight into the motivation for using entropy can be obtained by examining equation [3.8] which is the sum of two components. The first , $f^{*}(p)$, will be zero (its maximum value) at $p = p^{*}$. This function diminishes as p diverges from the p^{*} and, equivalently, as b diverges from the b^{*}. The second , $g^{*}(w)$ will be zero (its maximum) where each $w_k = .5$ corresponding to where errors (v_k) [3.4] are zero. However, the restriction [**??**] prevents both functions simultaneously achieving their maximums (at zero) unless the priors are fully consistent with the data. Therefore, the maximisation of [3.8] requires a trade-off between divergence of p^{*} from p (or equivalently b^{*} from b), and the divergence of γ from $M\,b$.

Derivatives and Optimisation

The first order derivatives for each of the functions $f(.)$ and $g(.)$ are:

$$\begin{aligned} f'_k(b) &= \sum_{j=1}^{K} \ln\left(\frac{s + v_j(b)}{s - v_j(b)}\right) \frac{m_{j,k}}{2s} \\ g'_k(b) &= -\ln\left(\frac{(b_k - \mathbf{b}_{l,k})(\mathbf{b}_{u,k} - b^{*}_k)}{(\mathbf{b}_{u,k} - b_k)(b^{*}_k - \mathbf{b}_{l,k})}\right) \frac{1}{\mathbf{b}_{u,k} - \mathbf{b}_{l,k}} \end{aligned} \tag{3.11}$$

and consequently

$$E'_k(b) = f'_k(b) + g'_k(b). \tag{3.12}$$

The second order derivatives are ($d_{k,i} = 1$, k=i, and zero otherwise)

$$\begin{aligned} f''_{k,i}(b) &= -\sum_{j=1}^{K} \left(\frac{1}{s^2 - v_j(b)^2}\right) m_{j,i} m_{j,k} \\ g''_{k,i}(b) &= -\frac{d_{k,i}}{(b_k - \mathbf{b}_{l,k})(\mathbf{b}_{u,k} - b_k)} \end{aligned} \tag{3.13}$$

and consequently,

$$E''_{k,i} = f''_{k,i}(b) + g''_{k,i}(b). \tag{3.14}$$

It is useful to view these quantities in vector and matrix form. The gradient and Hessian matrix for $g(b)$ are:

$$\begin{aligned} \bigtriangledown g(b)' &= \left(g'_1(b), \ldots\ldots\ldots\ldots\ldots g'_K(b)\right) \\ \bigtriangledown^2 g(b) &= \left\{g''_{k,i}(b)\right\}_{k,i}. \end{aligned} \tag{3.15}$$

The gradient vectors of $f(.)$ can be defined by using:

$$\begin{aligned} \alpha(b)' &= \left(\ln\left(\frac{s+v_1(b)}{s-v_1(b)}\right), ..., \ln\left(\frac{s+v_K(b)}{s-v_K(b)}\right)\right) \\ \Theta(b) &= \{\theta_{ij}\}\ \theta_{jj} = \frac{-1}{s^2 - v_j(b)^2}\ , \theta_{ij} = 0 \text{ otherwise.} \end{aligned} \tag{3.16}$$

Therefore,

$$\bigtriangledown f(b) = \frac{1}{2s} M' \alpha(b). \tag{3.17}$$

The Hessian matrix for $f(.)$ is

$$\bigtriangledown^2 f(b) = M' \Theta(b) M. \tag{3.18}$$

Therefore, the gradient vector for cross-entropy is

$$\bigtriangledown E(b) = \bigtriangledown f(b) + \bigtriangledown g(b) \tag{3.19}$$

and the Hessian for $E(b)$ is:

$$\bigtriangledown^2 E(b) = \bigtriangledown^2 f(b) + \bigtriangledown^2 g(b) \tag{3.20}$$

The problem at hand can therefore be represented as maximising E

$$\tilde{b} = \arg\max_b [E(b)] \tag{3.21}$$

By using these formulae, the function $E(b)$ can be maximised using a Gauss-Newton algorithm, should it be well behaved.

4. Moment Forms and Regression Equations

This section sets out some sufficient conditions under which the entropy estimates would converge in distribution to $\hat{b} = M^{-1}\gamma$ as the sample size grows.

Conditions

As before, let $\gamma = M\beta + v$, and assume the following conditions:

C1: $\beta \in (\mathbf{b}_l + \varphi\mathbf{1}, \mathbf{b}_u - \varphi\mathbf{1})$ for some small positive number φ and $\mathbf{1}$ is a conformable vector of ones

C2: M is constructed from a set of data with sample size T.

$$M = GN^{-1} \tag{4.1}$$

where

C2.1: G is the Cholesky decomposition of a $(K \times K)$ positive definite matrix $Q = G'G$ where Q converges in distribution to a positive definite matrix $Q^* = G^{*'}G^*$ as $T \to \infty$.

C2.2: N is a $(K \times K)$ matrix with negative powers of T in its diagonal

$$\begin{aligned} N &= \{n_{ij}\},\ n_{ii} = T^{-\varphi_i} \text{ and } n_{ij} = 0 \text{ otherwise} \\ \varphi_i &> 0 \text{ where } i = 1, ...K, \text{ and } \varphi_i \text{ is a real number} \end{aligned} \tag{4.2}$$

C3: v is a random vector which converges in distribution to a vector v^* with standard normal distribution with zero mean and an identity covariance matrix (where $\xrightarrow{d}$ denotes convergence in distribution, and $\mathcal{M}^{\mathcal{N}}(0, I)$ denotes a multivariate normal distribution with mean zero and identity covariance matrix)

$$v \xrightarrow{d} v^* \sim \mathcal{M}^{\mathcal{N}}(0, I) \tag{4.3}$$

C4: In addition to C2 and C3,

$$G^{-1}v \xrightarrow{d} G^{*-1}v^* \tag{4.4}$$

C5: s is a constant or increases with the sample size

$$s = s_0 T^{\lambda} \tag{4.5}$$

where s_0, λ are constants s.t. $\lambda \geq 0$, $s_0 > 0$ and

$$T^{2\lambda} N \to 0 \text{ as T} \to \infty \tag{4.6}$$

It follows that using the definition $\hat{b} = M^{-1}\gamma$, in conjunction with C2 and [3.1] gives

$$N^{-1}\left(\hat{b} - \beta\right) = G^{-1}v.$$

Therefore, under (C4)

$$N^{-1}\left(\hat{b} - \beta\right) \xrightarrow{d} G^{*-1}v^* \tag{4.7}$$

and consequently

$$\left(\hat{b} - \beta\right) = N.G^{-1}v \xrightarrow{d} 0. \tag{4.8}$$

Weak convergence to a degenerate distribution implies $\hat{b} \xrightarrow{p} b$, therefore under C1. Therefore the following theorem is now stated

Theorem 1 : Under C1-C5 the estimator $\hat{b} = M^{-1}\gamma$ and the cross-entropy estimator $\tilde{b}$ have the property

$$N^{-1}\left(\tilde{b} - \hat{b}\right) \xrightarrow{d} 0 \tag{4.9}$$

The proof of theorem 1 is given in the appendix.

Under Theorem 1, and [4.7] it follows that

$$GN^{-1}\left(\tilde{b} - \beta\right) \xrightarrow{d} v^* \tag{4.10}$$

5. Cointegrated Systems

This section considers the case where all the explanatory variables are integrated of order 1 ($I(1)$), and the errors are stationary. The theory can be extended to the case where deterministic variables are included also. However, the exposition is considerably simplified by the exclusion of these components. Again, assume the system in [**??**]. Additionally, denote the vector of residuals (u_t) and innovations in the regressors (e_t) as

$$\begin{aligned} \eta_t' &= \left(u_t' : e_t'\right) \\ e_t' &= \Delta x_t' - E\left(\Delta x_t'\right) . \end{aligned} \tag{5.1}$$

Here the following assumptions are made (the conditions under which these assumptions hold are outlined in a number of articles, see Phillips and Hansen, 1990):

A1**:** The vector η_t' is weakly stationary and obeys the invariance principle

$$T^{-\frac{1}{2}} \sum_{t=1}^{[rT]} \eta_t \xrightarrow{d} \omega_\eta(r) \tag{5.2}$$

where $\omega_\eta(r)$ is a vector Brownian Motion and is partitioned in accordance with the dimensions of u and e as

$$\omega_\eta'(r) = \underset{1\times k_y}{\left(\omega_u'(r)\right.} : \underset{1\times k_x}{\left.\omega_e'(r)\right)} . \tag{5.3}$$

The long-run covariance matrix of $\omega_\eta(r)$ can be defined as

$$E\left(\omega_\eta(1)\,\omega_\eta(1)'\right) = \Omega = \sum_{i=-\infty}^{\infty} E(\eta_0.\eta_i') = \begin{pmatrix} \Omega_{uu} & \Omega_{ue} \\ \Omega_{eu} & \Omega_{ee} \end{pmatrix} = \begin{pmatrix} \Omega_{u\eta} \\ \Omega_{e\eta} \end{pmatrix} . \tag{5.4}$$

The 'one-sided' long run covariance matrices are defined as

$$\Delta = \sum_{i=0}^{\infty} E(\eta_0 . \eta_i') = \begin{pmatrix} \Delta_{uu} & \Delta_{ue} \\ \Delta_{eu} & \Delta_{ee} \end{pmatrix} = \begin{pmatrix} \Delta_{u\eta} \\ \Delta_{e\eta} \end{pmatrix} . \tag{5.5}$$

A2: Ω is full rank

The following matrices are then constructed.

$$\begin{aligned} \underset{k_y \times (k_y + k_x)}{\kappa} &= (I_m : -\Gamma) \\ \underset{k_y \times k_x}{\Gamma} &= \Omega_{ue} \Omega_{ee}^{-1} \end{aligned} \tag{5.6}$$

and

$$\Omega_{**} = \kappa \Omega \kappa' . \tag{5.7}$$

Now define the moment equations as

$$\gamma = M\beta + v \tag{5.8}$$

where $N = IT^{-1}$ and

$$\begin{aligned} Q &= N \sum z_t \Omega_{**}^{-1} z_t' N = G'G \\ M &= GN^{-1} \\ \gamma &= G'^{-1} \left(N \sum z_t \Omega_{**}^{-1} \kappa \begin{pmatrix} y_t \\ e_t \end{pmatrix} - Vec\left(\Delta_{e\eta} \kappa' \Omega_{**}^{-1} \right) \right) \\ \kappa &= \left(I : -\Omega_{eu} \Omega_{ee}^{-1} \right) . \end{aligned} \tag{5.9}$$

The estimator $\hat{b} = M^{-1}\gamma$ is simply FM-SUR estimator based on the estimator of Phillips and Hansen (1990). This estimator is also outlined in Moon (1999) and Balcombe and Tiffin (2001), where the notation in the latter article has been adopted here. These articles establish that under conditions $A1$ and $A2$ Q weakly converges to a random matrix Q^* (see the Subsection 9 in the appendix for details) with a Cholesky decomposition $Q^* = G^{*\prime} G^*$:

$$N\left(\hat{b} - \beta \right) \xrightarrow{d} \mathcal{M}^{\mathcal{N}} \left(0, Q^* \right) . \tag{5.10}$$

With some straight forward algebra, it is evident that

$$v = G'^{-1} \left(N \sum z_t \Omega_{**}^{-1} \kappa \begin{pmatrix} u_t \\ e_t \end{pmatrix} - Vec\left(\Delta_{e\eta} \kappa' \Omega_{**}^{-1} \right) \right) \tag{5.11}$$

where (Appendix, Subsection 9).

$$N\sum z_t\Omega_{**}^{-1}\kappa\left(\begin{array}{c} u_t \\ e_t \end{array}\right) \xrightarrow{d} \mathcal{M}^{\mathcal{N}}\left(Vec\left(\Delta_{e\eta}\kappa'\Omega_{**}^{-1}\right), G^{*\prime}G^{*}\right) . \tag{5.12}$$

Therefore, under $A1$ and $A2$

$$G'^{-1}N\sum z_t\Omega_{**}^{-1}\kappa\left(\begin{array}{c} u_t \\ e_t \end{array}\right) \xrightarrow{d} \mathcal{M}^{\mathcal{N}}\left(G^{*\prime-1}Vec\left(\Delta_{e\eta}\kappa'\Omega_{**}^{-1}\right), I\right) \tag{5.13}$$

and, therefore, under $A1$ and $A2$

$$v \xrightarrow{d} v^{*} \equiv \mathcal{M}^{\mathcal{N}}\left(0, I\right) . \tag{5.14}$$

The estimator which maximises cross-entropy for the moment equations above, will be hereon referred to as MEFM-SUR, or MEFM for the single equation case. These moment equations obey the conditions C1-C4 in the previous section. Therefore, from Theorem 1,

$$N\left(\tilde{b}-\hat{b}\right) \xrightarrow{d} 0 . \tag{5.15}$$

Thus, the asymptotics relating to the FM-SUR estimator also extend to the entropy estimates of the cointegrated system.

Long-Run Covariance Estimation

As in the SUR case, the long-run covariance matrices must be estimated. This can be done in using the estimated residuals from the first round OLS estimates from OLS or using the entropy approximation and the long-run covariance matrices, then estimated using the procedures as in Andrews, (1991). The improved efficiency of the entropy estimates should also be reflected in improved estimates of the long-run matrices, and potentially improved bias correction, and inference.

The algorithms used here proceed by using iterated FM estimation in the first instance. The moment forms are then reconstructed and entropy is then maximised. The long-run matrices are then reconstructed and entropy is again maximised. This continues until there is no change in the parameters (within tolerance). As previously remarked, the use of estimated long-run covariances in the construction of the fully modified regressions will not be innocuous in small samples, since the estimated error will be

$$\hat{v} = \hat{G}'^{-1}\left(N\sum z_t\hat{\Omega}_{**}^{-1}\kappa\left(\begin{array}{c} u_t \\ e_t \end{array}\right) - Vec\left(\hat{\Delta}_{e\eta}\hat{\kappa}'\hat{\Omega}_{**}^{-1}\right)\right) .$$

As with the SUR case, s is set to 5, since it was found that setting $s = 3$ often generated errors outside their supports.

6. A Monte Carlo Study

In this section the entropy method is explored using Monte-Carlo methods. It shows how information concerning the approximate values of parameters can aid their estimation when using the entropy approaches outlined in the previous sections. It uses the example of a production function. However, the principles used here can be applied to any linear cointegrated model.

A simple Cobb-Douglas production function [6.1] is used for the following Monte Carlo study:

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t \tag{6.1}$$

where y_t is logged output, and the $x'_{it}s$ are logged input levels. The recent literature in production economics has focused on the use of alternative functional forms and indirect approaches to the estimation of production technology. However, arguably, it has paid little attention to developments within time series econometrics. The estimation of flexible forms is attractive, however, little is known about the properties of estimators containing quadratic terms or non-linear parameters when the data contains stochastic trends. For this reason there is a powerful argument for returning to simpler linear models.

In this study the inputs $x_{i,t}$ are treated as being (potentially) integrated of order one. The 'shocks' to production (u_t) may be due to stationary factors such as breakdowns, weather, transient changes in technology as well as technical change of a non-stationary nature. However, for the purposes of the Monte Carlo study these will treated as stationary. The transient components may alter the level of factors that are employed in a given time period so that *a priori* it is difficult to assert that u_t is strictly exogenous or serially uncorrelated.

In these circumstances FM estimation would be an appropriate estimation procedure. Cobb-Douglas production functions are commonly estimated using between 30 and 50 years of data with 3 or more inputs. Characteristically, production functions are assumed to have diminishing but positive marginal returns. Providing the main inputs into the production processes have been included, constant returns to scale might be considered a reasonable approximation.. However, usually researchers would expect deviations from constant returns to scale also, and may not wish to enforce this prior sharply. Therefore, in the absence of any other information, a reasonable prior expectation would be

$$\beta_1 = \beta_2 = ...\beta_k = \frac{1}{k} \tag{6.2}$$

Some inputs may *a priori* be thought to have higher marginal returns than others. Naturally, if researchers feel that they have better prior information than this, then they

may shift their priors accordingly. Another potential prior which might be explored is:

$$\beta_1 = \beta_2 = ...\beta_k = 0 \tag{6.3}$$

In this case the variables are extraneous. It should be recalled that position of the supports will also have an effect. Therefore, if the supports are centered above zero, then this prior will tend to offset the tendency of the supports to overstate the value of the parameters in the case where the variables are in fact extraneous. On the other hand, if the parameters diverge from this value to a large extent, then clearly this prior will induce poor performance in finite samples.

The following Monte Carlo Study examines the performance of the, OLS, MEOLS FM, and MEFM procedures outlined above. It generates I(1) regressors and incorporates some moderate serial correlation and endogeneity between the innovations in the regressors and the error. The introduction of serial correlation and endogeneity that the Fully Modified estimator should perform better than OLS, at least in large samples. Without the serial correlation and/or endogeneity OLS will dominate FM, since FM has no potential advantages in this case.

Simply generating data which conformed exactly to the expectational priors (as in [6.2]) could give a falsely positive reflection on the entropy procedures. Therefore, the data was generated so as to loosely conform to these priors, but also in many cases they differ substantially. The Monte-Carlo design generates data where the priors (as in the expected values) will be correct on average (where $\beta_i^* = \frac{1}{k}$) but incorrect in any given trial. However, the impact of setting the prior expectations to $\beta_i^* = 0$ is also explored. In the case where $k = 1$, this prior is severely downwardly biased for the generating process. However, it might be useful for readers to be able evaluate the impact of this false prior. The design is therefore as follows:

- For a given k,
  - q is generated as a $k \times 1$vector of uniformly distributed variables between 0 and 1
  - z is generated as a standard normal random variable
  - $\alpha' = (a_1, ..., a_k)$ (the parameter vector in the Cobb Douglas equation) is generated using
$$\alpha = \frac{q'}{\sum_{i=1}^{k} q_i} (1 + .125z)$$
  - x_t are generated as I(1) processes, $t = 1, ...T$ by first generating the inno-

vations using

$$\begin{aligned}\begin{pmatrix} u_t \\ e_t \end{pmatrix} &= .5\Phi \begin{pmatrix} u_{t-1} \\ e_{t-1} \end{pmatrix} + \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix} \\ \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix} &\sim \mathcal{M}^{\mathcal{N}}(0, \Phi) \\ \Phi &= \begin{pmatrix} I & .25\mathbf{1}_k' \\ .25\mathbf{1}_k & I \end{pmatrix}\end{aligned}$$

with $x_t = \sum_{i=1}^{t} e_i$; and,

– y_t is generated as

$$y_t = x_t'\alpha + u_t.$$

This was repeated 2500 times for each k (1 and 4) and T=30, 50 100 and 1000. k was set to two values, so as to get an idea of the impact of dimension on the performance of the estimators. *k=4* was chosen so as to correspond with the empirical example given latter on in this section which uses four inputs. Each pseudo sample was then estimated using OLS, MEOLS, FM and MEFM. The results for these experiments are given in Tables 1 and 2.

The correlation structure used in this study is a similar to studies such as Xiao and Phillips (2002) except that it incorporates correlations between the innovations and the errors. Other designs were also used including moving average serial correlation, and no serial correlation at all. These alternative designs did not change the broad conclusions that are made *vis-a-vis* the performance of the entropy procedures relative to their 'standard' counterparts. The results for these are not given here, since they paint broadly the same picture as the results which are subsequently presented. Alternative designs do effect the relative performance of FM relative to OLS. However, it is not the aim of this paper to cover this topic, which has already been the subject of extensive Monte-Carlo trials (for example Phillips and Hansen, 1990, Haug, 1999). Denoting the estimated and actual parameters from each trial are denoted as $\hat{\beta}_n$ and α_n respectively, the average root mean square of the estimated elasticities was produced as in [6.4]

TABLE 1: One Input Variable (k=1)

		T=30	T=50	T=100	T=1000
OLS	ARMSE	.073	.044	.022	.0021
	(E-size 0.10)	(.286)	(.307)	(.324)	(.336)
FM	ARMSE	.067	.040	.019	.0018
	(E-size 0.10)	(.248)	(.218)	(.179)	(.127)
MEOLS	ARMSE	.059	.040	.021	.0021
$\beta^* = 1$	(E-size 0.10)	(.212)	(.283)	(.311)	(.336)
MEFM	ARMSE	.056	.038	.018	.0018
$\beta^* = 1$	(E-size 0.10)	(.178)	(.199)	(.176)	(.126)
MEOLS	ARMSE	.071	.043	.021	.0021
$\beta^* = 0$	(E-size 0.10)	(.288)	(.331)	(.327)	(.320)
MEFM	ARMSE	.089	.044	.019	.0019
$\beta^* = 0$	(E-size 0.10)	(.315)	(.255)	(.182)	(.125)

$$ARMSE = \frac{1}{kN}\sum_{n=1}^{N}\sqrt{\sum_{j=1}^{k}\left(\hat{\beta}_{n,j} - \alpha_{n,j}\right)^2} \tag{6.4}$$

The supports were set so as to include the generated parameters and the expectations were set as $\frac{1}{k}$ for one set of trials, and 0 for another set of trials. The intervals for the intercept were set to be 'non-informative' (very wide) $\beta_0 \in (-10000, +10000)$. The supports and the expected values for the elasticities were set at

$$\begin{aligned} \beta_j &\in \left(-\frac{1}{k}, 1+\frac{1}{k}\right) \text{ j=1,...k} \\ \beta_j^* &= \frac{1}{k} \\ &\quad or \\ \beta_j^* &= 0 \end{aligned}$$

Therefore, the supports would become close to (0,1) as k increases. s was set to five, for reasons discussed earlier in the paper.

The number of rejections at the 10% nominal level is summarised also in Tables 1 and 2 (Denoted E-Size 0.10) using an F-test for the joint restrictions that $\beta_{j,n} = \alpha_{j,n}$ for all j in each trial. The F-statistics for the k restrictions are constructed as by dividing

the conventional Wald statistic by the number of restrictions. This was then treated as an $F(k, T-k-1)$ distribution. This will have less of a tendency than the Wald to over-reject in finite samples.

TABLE 2: Four Input Variables (k=4)

		T=30	T=50	T=100	T=1000
OLS	ARMSE	.106	.069	.037	.004
	(E-size, 0.10)	(.666)	(.770)	(.786)	(.825)
FM	ARMSE	.114	.055	.020	.0016
	(E-size, 0.10)	(.780)	(.713)	(.498)	(.152)
MEOLS $\beta_i^* = .25$	ARMSE	.067	.051	.033	.004
	(E-size 0.10)	(.345)	(.598)	(.752)	(.826)
MEFM $\beta_i^* = .25$	ARMSE	.063	.043	.019	.0016
	(E-size 0.10)	(.483)	(.590)	(.462)	(.156)
MEOLS $\beta_i^* = 0$	ARMSE	.063	.046	.028	.003
	(E-size 0.10)	(.464)	(.663)	(.780)	(.846)
MEFM $\beta_i^* = 0$	ARMSE	.070	.040	.016	.0015
	(E-size 0.10)	(.652)	(.662)	(.496)	(.154)

The results for the trials in Tables 1 and 2 indicate that:

- The efficiency gains (as measured by the reduction in ARMSE) for FM regression over OLS can be substantial. In all but one case (T=30, k=4) FM estimation improves on OLS. This is expected, however, it also illustrates that for FM to be more efficient than OLS (even with endogeneity and serial correlation) there must be reasonably large sample sizes and/or few parameters;
- Where the priors are set at 1, and .25 for k=1 and k=4 respectively:
  - The entropy procedures significantly improve on both the OLS and the FM procedures in terms of efficiency except at very large sample sizes where the entropy estimates become identical to the non-entropy estimates (consistent with the theory);
  - The MEFM has the lowest ARMSE in all examples where T<1000. Even when OLS dominated the FM estimates, the MEFM estimates dominated the MEOLS at T=30 and k=4;

- The entropy procedures mitigate, but do not remove, the tendency for the (adjusted) F-tests to over-reject as indicated by the empirical size in the tables. Consistent with previous work (for example Xaio and Phillips(2002)), Wald tests have biases that decrease with sample size but increase with the dimension of the tests, which the degrees of freedom adjustments used by the F-test do little to decrease. In the case where k=4, these biases were significant, even where T=1000. The large empirical size of these tests are partly due to the residual second order bias in the test statistics, but possibly also due to the fact that the standard errors are understated. Therefore, while the entropy approaches probably give better estimates of the standard errors, these standard errors may still be understated in finite samples.

- Where the priors are set to zero ($\beta_j^* = 0$)

  - Where $k = 1$, the results are, as expected, slightly worse than for the non-entropy case, and the entropy where using unbiased priors. The performance for the entropy case and the non-entropy methods become virtually identical for T greater or equal to one hundred, and are only slightly worse for T=50.
  - Where k is four, the entropy results remain better than for their non-entropy counterparts. In certain instances they seem to do slightly better relative to the unbiased priors, but this is not uniformly the case. This suggests that the entropy is not overly sensitive to misspecified priors, and is dominated by the data relatively quickly.

In summary, the Monte Carlo results here indicate that both entropy can improve over both OLS and FM results when the prior information is informative, but also inexact. In very small sample sizes with many variables, the utility of using FM is likely to be minimal or negative. However, in many practical cases there may be additional advantages in using FM estimation in conjunction with the Entropy procedure.

7. Application to South African Agriculture

The following data uses chained divisa indices for inputs to South African Agriculture for 1947 until 1994 (inclusive T=48). The data is as in Thirtle *et.al* , (1993), though it has been revised and updated. Thirtle *et al.* also contains a description of the data. It contains four inputs and one aggregate output. The inputs are *Labour*, *Land*, *Intermediate Inputs* (e.g. seeds fertiliser and so forth) and *Capital*.

TABLE 3: Results for South African Data.

	ME OLS		OLS		ME FM		FM	
	β	$Se\,(\beta)$	β	$Se\,(\beta)$	β	$Se\,(\beta)$	β	$Se\,(\beta)$
Intercept	.602	2.46	3.447	2.26	.888	2.39	3.108	2.16
Time Trend	.013	.003	.0072	.002	.013	.003	.0067	.002
Labour	.207	.118	.2855	.109	.136	.115	.104	.104
Land	.218	.600	-0.567	.551	.219	.584	-.309	.527
Inter Inputs	.407	.081	.588	.075	.397	.079	.577	.071
Capital	.034	.056	-0.045	.051	.056	.054	-.034	.049
Test: CRS	[.806]		[.138]		[.718]		[.164]	

The values in square parentheses are the prob values for an F-test for constant returns to scale.

The logged variables have been tested for unit roots using a range of tests , both under the null of a unit root and under the null of stationarity (not presented). All are broadly consistent with I(1) behaviour with drift. *Labour* and *Land* have tended to drift downwards over the sample period, whereas *Intermediate Inputs* and *Capital* have shown significant increases over the period. The results for the production function are presented in Table 3. All variables are logged and a time trend is included which may 'soak up' any deterministic trends such as long-term technical progress. A test for cointegration using the Augmented Dickey-Fuller and Phillips-Perron tests on the OLS residuals gives a value of -5.033 and -5.042 respectively (no lags selected using the Dickey Fuller), which are less than their critical value (-4.49) at the 5% level of significance, indicating the rejection of 'no-cointegration'.

Turning to the OLS and FM results first which are contained in the second set and fourth set of results in Table 3, it can be observed that the results are very poor indeed and the coefficients for land and capital are negative. However, the standard errors for these coefficients are very large, and they are insignificantly different from zero. The only highly significant input according to both sets of results is *Intermediate Inputs*.

For the entropy results, the supports and expectations have been set as in the Monte Carlo trials (k=4). The time trend and the intercept have intervals set extremely wide so as to make the priors on these parameters diffuse. The introduction of the supports and expected values has resulted having no negative coefficients, but they still reflect

the data to a large degree. From Table 3 it appears that the use of MEFM has had little impact relative to MEOLS in this instance, since the elasticities are ostensibly similar as are the standard errors. In both cases the large standard errors underline that little confidence can be held in the precise values of the parameters in this production function.

The fact that entropy is not a panacea for inadequate sample information should not be used as an argument against entropy, or any other method which utilises prior information. Rather, it is the contention here that situation here has been transformed from one where the results were of little or no use, to one where some guarded inferences about the elasticities can be made. It seems, for instance, that variations in the level of *Capital* alone appear to have a relatively small impact on the level of output. Moreover, there is evidence that much of the increases in output are likely to be due to variations in the intermediate inputs, even in the long-run. While in all cases there is evidence for decreasing returns to scale, the tests for constant returns to scale in the last row of Table 3 suggest constant or increasing returns to scale cannot be rejected, even at very high levels of significance.

8. Conclusion

This paper has outlined how prior information can be integrated into estimates of parameters within cointegrating regressions using entropy. It showed that once in appropriate moment form, the cross-entropy estimate converged to the FM-SUR estimator at a rate which made it asymptotically equivalent to the FM-SUR estimator, providing the supports for the errors and parameters were sufficiently large. Given prior information on the values and supports of parameters, the entropy techniques have the potential to reduce the MSE of parameter estimates in both stationary and cointegrated systems.

The Monte Carlo evidence presented in the paper demonstrated that even when this prior information was inexact, it improved the efficiency of the estimates, and reduced the bias in the standard errors. However, while the poor performance of F-tests were mitigated using the entropy approach, these tests continued to over-reject to a large extent, even when complemented with prior information.

TECHNICAL APPENDIX

Notation

The notation $|b|$ where b is a vector in $\mathbb{R}^K$ denotes the vector of absolute values of that vector whereas $\|b\|$ denotes the Euclidean length of b. The inequality between two vectors, such as $b < v$, indicates that every element of b is less than v, and $\max(b)$ denotes the maximum element of the vector b. An open (closed) ball of radius ε in $\mathbb{R}^K$ around a point b is denoted as $S(b,\varepsilon)$ ($S[b,\varepsilon]$). (That is, $S(b,\varepsilon)$ is the set of all vectors x for which $\|x-b\| < \varepsilon$; where $b, x \in \mathbb{R}^K$ and the closed ball is defined in the same way with $\|x-b\| \leq \varepsilon$). If $\mathbb{B}_{f,T}$ denotes an open set, then $\mathbb{B}^c_{f,T}$ denotes the closure of $\mathbb{B}_{f,T}$ (as in 3.19 Apostol, 1974). **1** denotes a conformable vector of ones. All other quantities are as defined in the main text.

The negative cross-entropy function maximised in the paper (3.10)

$$E_T(b) = f_T(b) + g(b) \tag{8.1}$$

is the sum of the two entropy functions. The first is:

$$g(b) = -\sum_{k=1}^{K} p_k(b) \ln\left(\frac{p_k(b)}{p_k^*}\right) - \sum_{k=1}^{K} (1-p_k(b)) \ln\left(\frac{1-p_k(b)}{1-p_k^*}\right) \tag{8.2}$$

where

$$p_k(b) = \left(\frac{b_k - \mathbf{b}_{l,k}}{\mathbf{b}_{u,k} - \mathbf{b}_{l,k}}\right) \quad and \quad p_k^* = \left(\frac{b_k^* - \mathbf{b}_{l,k}}{\mathbf{b}_{u,k} - \mathbf{b}_{l,k}}\right). \tag{8.3}$$

The second is

$$f_T(b) = -\sum_{k=1}^{K} w_k(b) \ln(w_k(b)) - \sum_{k=1}^{K} (1-w_k(b)) \ln(1-w_k(b)) \tag{8.4}$$

where

$$w_k(b) = \left(\frac{v_k + s}{2s}\right) = \frac{\gamma_k - m_k' b + s}{2s}. \tag{8.5}$$

$g(.)$ is non-stochastic function which only depends on a $K \times 1$ vector b, whereas $f_T(.)$ is stochastic sequence of functions since γ and M are stochastic. For this reason it is useful to subscript $f(.)$ with T. Consequently, $E_T(.)$, $f_T(.)$ and their domains usefully acquire T subscripts here, although not in the main text. The following can be verified straightforwardly:

a) The domain of $g(.)$ is $\mathbb{B}_g = \{b : \mathbf{b}_u < b < \mathbf{b}_l\}$;

b) The codomain of $g(.)$ is $\mathbb{G} = (-\infty, g(b^*)]$ where $g(b^*) = 0$;

c) The domain of $f_T(.)$ is $\mathbb{B}_{f,T} = \{b : |v_k(b)| < s\ ,\ k = 1, 2, ..K\}$;

d) The codomain of $f_T(.)$ is $\mathbb{F} = \left(-\infty, f_T\left(\hat{b}\right)\right]$ where $f_T\left(\hat{b}\right) = -K \ln\left(\frac{1}{2}\right)$ (since at $\hat{b}$, $\hat{v} = 0$ and $w_k = \frac{1}{2}$); and,

e) The domain of $E_T(.)$ is $\mathbb{B}_{E,T} = \mathbb{B}_{f,T} \cap \mathbb{B}_g$ and its Codomain $\mathbb{E}_T$ is bounded from above (for all T) by $-K \ln \frac{1}{2}$ (the sum of the maximums of $f_T(.)$ and $g(.)$)

Lemma 1: f_T and g are finitely twice continuously-differentiable ($w.r.t$ to b) everywhere within their domains for all T.

Proof of Lemma 1: $f^*(.)$ and $g^*(.)$ (defined in [3.9]) are differentiable on $I^K = (0,1) \times (0,1) ... \times (0,1)$ and $p(b)$ and $w(b)$ [3.6] [3.7] are differentiable with respect to b_k at any point in $\mathbb{R}^K$. Therefore, for any value b for which $p(b) \in I^K$, $w(b) \in I^K$ the derivatives of the composite functions of f_T and g must (Theorem 5.5. Apostol, 1974) exist. Applying the chain rule, the partial derivatives are

$$\begin{aligned} f'_{k,T}(b) &= \sum_{j=1}^{K} \ln\left(\frac{s + v_j(b)}{s - v_j(b)}\right) \frac{m_{j,k}}{2s} \qquad (8.6) \\ g'_k(b) &= -\ln\left(\frac{(b_k - \mathbf{b}_{l,k})(\mathbf{b}_{u,k} - b_k^*)}{(\mathbf{b}_{u,k} - b_k)(b_k^* - \mathbf{b}_{l,k})}\right) \frac{1}{\mathbf{b}_{u,k} - \mathbf{b}_{l,k}}. \end{aligned}$$

For $\mathbb{B}_{f,T} = \{b : |v_k(b)| < s\ ,\ k = 1, 2, ..K\}$, $f'_{k,T}(b)$ is continuously defined. The condition that $\{b : \mathbf{b}_u < b < \mathbf{b}_l\}$, implies that $g'_k(b)$, is continuously defined. The first order derivatives above are composites of continuous differentiable functions on the domains of $g(b)$ and $f_T(b)$ respectively. Therefore, the second order derivatives are ($for \quad d_{k,i} = 1$, $k = i$, and zero otherwise)

$$\begin{aligned} f''_{k,i,T}(b) &= -\sum_{j=1}^{K} \left(\frac{1}{s^2 - v_j(b)^2}\right) m_{j,i} m_{j,k} \qquad (8.7) \\ g''_{k,i}(b) &= -\frac{d_{k,i}}{(b_k - \mathbf{b}_{l,k})(\mathbf{b}_{u,k} - b_k)}. \end{aligned}$$

$f''_{k,i,T}(b)$ is therefore defined providing for each j, $|v_j(b)| < s$ ($i.e.$ $\mathbb{B}_{f,T}$) and the domain of $g''_{k,i}(b)$ is defined on everywhere on $\mathbb{R}^K$ except at the boundary of $\mathbb{B}_g$.

Lemma 2: Negative Cross Entropy is a concave function *w.r.t.* b everywhere on its domain.

Proof of Lemma 2:

Under Lemma 1, the condition that the Hessian Matrices for $f_T()$ and $g(.)$ are negative definite is sufficient for concavity, (Magnus and Neudecker, 1994, Theorem, 7, note 2).

The Hessian for $g(b)$ is

$$\begin{aligned} \bigtriangledown^2 g(b) &= \left\{ g''_{k,i}(b) \right\}_{k,i} = \left\{ -\frac{d_{k,i}}{(b_k - \mathbf{b}_{l,k})(\mathbf{b}_{u,k} - b_k)} \right\}_{k,i} \\ d_{k,i} &= 1 \text{ where i=k, and 0 otherwise.} \end{aligned} \tag{8.8}$$

which is a diagonal matrix with diagonal negative elements (and therefore negative definite). The Hessian matrix for $f_T()$ is

$$\bigtriangledown^2 f_T(b) = M' \Theta(b) M \tag{8.9}$$

where the center matrix $\Theta(b)$ is also diagonal with diagonal negative elements

$$\Theta(b) = \{\theta_{ij}\} \;\; \theta_{jj} = \frac{-1}{s^2 - v_j(b)^2} \;, \theta_{ij} = 0 \text{ otherwise}. \tag{8.10}$$

Since M is invertible, for any non-zero vector z, $z' \bigtriangledown^2 f(b) z = z' M' \Theta(b) M z = y' \Theta(b) y < 0$. Noting that the sum of two convex (concave) functions is also convex (concave) (Berck and Sydsaeter, 12.10) completes the proof.

Lemma 3:
If $\mathbb{B}^c_{f,T} \subset \mathbb{B}_g$, then $g(.)$ is bounded (above and below) on $\mathbb{B}_{f,T}$.

Proof of Lemma 3: If $\mathbb{B}^c_{f,T} \subset \mathbb{B}_g$ then $g(.)$ is defined on $\mathbb{B}^c_{f,T}$ and since $\mathbb{B}^c_{f,T}$ is a compact set, and $g(.)$ is continuous on $\mathbb{B}^c_{f,T}$ then $g(.)$ is bounded on $\mathbb{B}^c_{f,T}$ (Apostol, 1974 Theorem 4.25). Therefore, $g(.)$ has a finite supremum and infimum on $\mathbb{B}_{f,T}$.

Lemma 4
If $\mathbb{B}^c_{f,T} \subset \mathbb{B}_g$, then $E_T(.)$ is defined on $\mathbb{B}_{f,T}$ and $\mathbb{B}_{f,T}$ contains a maximum point $(\tilde{b})$ at which $\nabla E_T\left(\tilde{b}\right) = 0$

Proof of Lemma 4: From Lemma 1, $g(.)$ has a finite infimum and supremum on $\mathbb{B}_{f,T}$. Consequently,

$$-Kln\left(\frac{1}{2}\right)+\sup_{\mathbb{B}_{f,T}}g(b)\geq\sup\left(E_T(b)\right)\geq-Kln\left(\frac{1}{2}\right)+\inf_{\mathbb{B}_{f,T}}g(b)\ .$$

As b approaches the boundary of $\mathbb{B}_{f,T}$ from any direction, $f_T(b)\to-\infty$. Consequently, $E_T(.)\to-\infty$ as b approaches its boundary from any direction since $g(b)$ is bounded above and below. Therefore, a point can always be chosen sufficiently close to the boundary of $\mathbb{B}_{f,T}$ for which $E_T(.)$ is less than $\sup\left(E_T(b)\right)$. The supremum must therefore be contained within $\mathbb{B}_{f,T}$ and must therefore be a maximum. The second part of the Lemma $\left(\nabla E_T\left(\tilde{b}\right)=0\right)$ follows from the fact that under Lemma 1, the derivatives of E_T are finite over the $\mathbb{B}_{f,T}$ (though not bounded). Using Apostol (1974) p.362, Ex 12, if E_T contains a maximum within $\mathbb{B}_{f,T}$ then the existence of finite partial derivatives within $\mathbb{B}_{f,T}$, is sufficient to ensure that the derivatives are zero at the maximum point.

Lemma 5

$$\lim_{T\to\infty}\text{Prob}\left(\mathbb{B}_{f,T}^{c}\subset\mathbb{B}_{g}\right)=1 \tag{8.11}$$

Proof of Lemma 5:

The proof of Lemma 5, is in two parts. The first part, shows that for any point that is a fixed distance from $\hat{b}$ will asymptotically **not** belong in $\mathbb{B}_{f,T}^{c}$ with probability one. Conversely, the second part shows that for any point within a radius of $\frac{\varphi}{2}$ from $\hat{b}$ will asymptotically be a member of $\mathbb{B}_{g}$ with probability one. Therefore, any point which is close enough to $\hat{b}$ to be a member of $\mathbb{B}_{f,T}^{c}$ must asymptotically also be a member of $\mathbb{B}_{g}$ with probability one.

Part 1:

For any point (b) (using notation defined at the beginning of Section 3)

$$b=\hat{b}-M^{-1}v(b)\ . \tag{8.12}$$

An open K-Ball around $\hat{b}$ can be expressed as:

$$S(\hat{b},\varepsilon)=\left\{b:\left(b-\hat{b}\right)'\left(b-\hat{b}\right)=v(b)'M^{-2}v(b)<\varepsilon^{2}\right\}\ . \tag{8.13}$$

The Closure of $\mathbb{B}_{f,T}$ is:

$$\mathbb{B}_{f,T}^{c} = \left\{ b : |v(b)| = \left| M\left(b - \hat{b}\right) \right| \leq \mathbf{1}s \right\}. \tag{8.14}$$

Define the vector

$$\left(b - \hat{b}\right)' N^{-1} = h(b)'. \tag{8.15}$$

Under C2 and C5, $M^{-2} = N^{-1}Q^{-1}N^{-1}$. Therefore, any point in $\mathbb{B}_{f,T}^{c}$ has the property that:

$$v(b)' v(b) = h(b)' Q^{-1} h(b) \leq K s_0^2 T^{2\lambda}. \tag{8.16}$$

The middle part of [8.16] can be decomposed as:

$$h(b)' Q^{-1} h(b) = h(b)' \left(Q^{-1} - N\right) h(b) + \left(b - \hat{b}\right)' N^{-1} \left(b - \hat{b}\right). \tag{8.17}$$

Under C2,

$$Q^{-1} - N \xrightarrow{p} Q^{*-1} \tag{8.18}$$

where Q^{*-1} is positive definite. Under C5, $T^{-2\lambda}N^{-1}$ diverges. Therefore, a small positive κ can exists for which $\left(b - \hat{b}\right)' T^{-2\lambda} N^{-1} \left(b - \hat{b}\right) > T^{\kappa} \left(b - \hat{b}\right)' \left(b - \hat{b}\right)$ and for $\left(b - \hat{b}\right)' \left(b - \hat{b}\right) > \varepsilon^2$,

$$\lim_{T \to \infty} \Pr ob\left(T^{\kappa} \left(b - \hat{b}\right)' \left(b - \hat{b}\right) \leq K s_0^2 \right) = 0. \tag{8.19}$$

Therefore, any point more than a fixed Euclidean distance $\varepsilon > 0$ from $\hat{b}$ will not asymptotically lie within $\mathbb{B}_{f,T}^{c}$ with probability one. Consequently,

$$\lim_{\beta \to \infty} \Pr ob\left(\mathbb{B}_{f,T}^{c} \subset S\left(\hat{b}, \varepsilon\right) \right) = 1. \tag{8.20}$$

Part 2:

Under C1, $S(\beta, \varphi) \subset \mathbb{B}_g$. Therefore, given the consistency of $\hat{b}$, for any $\varepsilon > 0$

$$\lim_{T \to \infty} \Pr ob\left(\left\| \hat{b} - \beta \right\| < \varepsilon \right) = 1 \tag{8.21}$$

C1-C5, in turn, implies that for any $\varepsilon \in \left(0, \frac{\varphi}{2}\right)$

$$\lim_{T \to \infty} \Pr ob\left(S\left(\hat{b}, \varepsilon\right) \subset S(\beta, \varphi) \subset \mathbb{B}_g \right) = 1. \tag{8.22}$$

Therefore

$$\lim_{T\to\infty} \text{Pro}b\left(\mathbb{B}^c_{f,T} \subset S\left(\hat{b},\varepsilon\right) \subset \mathbb{B}_g\right) = 1\,. \tag{8.23}$$

$\mathbb{B}_{E,T}$ will become non-empty (in probability) since it becomes equivalent to $\mathbb{B}_{f,T}$ which by definition is a non-empty K-Ball around $\hat{b}$

The following Lemmas are most easily stated and proved as a group.

Lemmas 6.1, 6.2: Under C1-C5:

$$\begin{aligned} &6.1)\ \bigtriangledown\, g\left(\hat{b}\right) \xrightarrow{d} \bigtriangledown g\left(\beta\right), \\ &6.2)\ \bigtriangledown^2\, g\left(\hat{b}\right) \xrightarrow{d} \bigtriangledown^2 g\left(\beta\right)\,. \end{aligned} \tag{8.24}$$

Proof of Lemmas 6.1 and 6.2:

From Lemma 1, $\bigtriangledown g\left(\beta\right)$ and $\bigtriangledown^2 g\left(\beta\right)$ exist and are finite . If cross entropy is defined at $\hat{b}$, by the continuous mapping theorem (Davidson, 1994, Theorem, 22.11 the consistency of $\hat{b}$, and Lemma 5, 6.1 and 6.2. hold.

Lemmas 7.1, 7.2: Under C1-C5:

$$\begin{aligned} &7.1)\ \ \bigtriangledown\, f_T\left(\hat{b}\right) = 0; \\ &7.2)\, T^{2\lambda} N\, \bigtriangledown^2\, f_T\left(\hat{b}\right) N = s_0^{-2} Q\,. \end{aligned} \tag{8.25}$$

Proof of Lemmas 7.1. and 7.2:

Lemma 7.1 is trivially proved by observing that $\alpha\left(\hat{b}\right) = 0$, and therefore

$$\bigtriangledown f_T\left(\hat{b}\right) = \frac{1}{2s} M^{'} \alpha\left(\hat{b}\right) = 0\,. \tag{8.26}$$

Lemma 7.2. follows from

$$\bigtriangledown^2 f_T\left(\hat{b}\right) = M'\Theta\left(\hat{b}\right)M = -\frac{1}{s^2}M'M\,. \tag{8.27}$$

From C2 and C5

$$\bigtriangledown^2 f_T\left(\hat{b}\right) = -\frac{1}{s^2}N^{-1}G'GN^{-1} = -T^{-2\lambda}s_0^{-2}N^{-1}QN^{-1}\,. \tag{8.28}$$

Therefore,

$$T^{2\lambda}N\bigtriangledown^2 f_T\left(\hat{b}\right)N = -s_0^{-2}Q\,. \tag{8.29}$$

Lemma 8.1 and 8.2: Under C1 to C5 (and defining two new quantities $W_1\left(\hat{b}\right)$ and $W_2\left(\hat{b}\right)$) :

$$\begin{aligned} 8.1 \quad &: \quad W_1\left(\hat{b}\right) = T^{2\lambda}N\bigtriangledown E_T\left(\hat{b}\right) \xrightarrow{d} 0; \\ 8.2 \quad &: \quad W_2\left(\hat{b}\right) = T^{2\lambda}N\bigtriangledown^2 E_T\left(\hat{b}\right)N \xrightarrow{d} -s_0^{-2}Q^*\,. \end{aligned}$$

Proof of Lemma 8.1:

From Lemma 7.1

$$W_1\left(\hat{b}\right) = T^{2\lambda}N\bigtriangledown E_T\left(\hat{b}\right) = T^{2\lambda}N\left(\bigtriangledown g\left(\hat{b}\right) + \bigtriangledown f_T\left(\hat{b}\right)\right) = T^{2\lambda}N\left(\bigtriangledown g\left(\hat{b}\right)\right)\,. \tag{8.30}$$

The second component

$$T^{2\lambda}N\bigtriangledown g\left(\hat{b}\right) \xrightarrow{d} 0 \tag{8.31}$$

follows from Lemma 6.1 and $T^{2\lambda}N \to 0$ (under C5).

Proof of Lemma 8.2:

Expanding $W_2\left(\hat{b}\right)$ and then using Lemma 7.2 :

$$\begin{aligned} W_2\left(\hat{b}\right) &= T^{2\lambda}N\bigtriangledown^2 f_T\left(\hat{b}\right)N + T^{2\lambda}N\bigtriangledown^2 g\left(\hat{b}\right)N \\ &= s_0^{-2}Q + T^{2\lambda}N\bigtriangledown^2 g\left(\hat{b}\right)N\,. \end{aligned} \tag{8.32}$$

From Lemma 6.2, $\bigtriangledown^2 g\left(\hat{b}\right) \xrightarrow{d} \bigtriangledown^2 g\left(\beta\right)$, and under C5, $T^{2\lambda}N \to 0$. Therefore:

$$T^{2\lambda}N\bigtriangledown^2 g\left(\hat{b}\right)N \xrightarrow{d} 0\,. \tag{8.33}$$

Proof of Theorem 1:

Theorem 1 claimed that under C1-C5 the estimator $\hat{b} = M^{-1}\gamma$ and the cross-entropy estimator $\tilde{b}$ have the property

$$N^{-1}\left(\tilde{b} - \hat{b}\right) \xrightarrow{d} 0\,. \tag{8.34}$$

Proof of Theorem 1:

Lemmas 1 through to 3 establish that if $\mathbb{B}^c_{f,T} \subset \mathbb{B}_g$ then cross entropy will be defined, the derivatives will exist, and negative cross entropy will have a maximum at a point where the derivatives are equal to zero. Lemma 5 establishes that $\mathbb{B}^c_{f,T} \subset \mathbb{B}_g$ will be met asymptotically with probability one. Therefore, Lemmas 1 through 5, establish that the cross-entropy estimator will exist on the interior of $\mathbb{B}_{f,T}$ asymptotically with probability one. The functions are concave everywhere on $\mathbb{B}_{f,T}$ and that the maximum will have a derivative of zero. Therefore, using an expansion for $\tilde{b}$ (the entropy estimate) around $\hat{b} = M^{-1}\gamma$

$$\nabla E\left(\tilde{b}\right) = 0 = \nabla E\left(\hat{b}\right) + \nabla^2 E\left(\hat{b}\right)\left(\tilde{b} - \hat{b}\right) + o\left(\tilde{b} - \hat{b}\right), \tag{8.35}$$

a manipulation of [8.35] gives

$$\begin{aligned} N^{-1}\left(\tilde{b} - \hat{b}\right) &= -N^{-1}\left(\nabla^2 E\left(\hat{b}\right)\right)^{-1}\nabla E\left(\hat{b}\right) \\ &\quad - \left(N\left(\nabla^2 E\left(\hat{b}\right)\right)NT^{2\lambda}\right)^{-1} NT^{2\lambda} o\left(\tilde{b} - \hat{b}\right). \end{aligned} \tag{8.36}$$

Using the definitions in Lemmas 8.1 and 8.2,

$$N^{-1}\left(\tilde{b} - \hat{b}\right) = -W_2\left(\hat{b}\right)^{-1} W_1\left(\hat{b}\right) - W_2\left(\hat{b}\right)^{-1} NT^{2\lambda} o\left(\tilde{b} - \hat{b}\right). \tag{8.37}$$

By using Lemmas 8.1, 8.2 and C5, each of the components on the right hand side converge to zero in distribution. Therefore,

$$N^{-1}\left(\tilde{b} - \hat{b}\right) \xrightarrow{d} 0. \tag{8.38}$$

which completes the proof of Theorem 1.

Remark.

Note that the above also suggests an approximate relationship between the entropy estimate $\left(\tilde{b}\right)$ and $\left(\hat{b}\right)$ as

$$\tilde{b} \approx \hat{b} - \left(\nabla^2 E\left(\hat{b}\right)\right)^{-1} \bigtriangledown g\left(\hat{b}\right) \tag{8.39}$$

which may be useful approximation in practice.

9. **Weak Convergence Results**

These results are outlined in the work of Phillips, for which Phillips, (1990) is a starting reference. Using similar notation to that in Balcombe and Tiffin (2001), Equation A1 and A2 give

$$N \sum Z_t.\Omega_{**}^{-1} Z_t' N = G'G \xrightarrow{d} \int W_e \Omega_{**}^{-1} W_e' = G^{*\prime} G^* \tag{8.40}$$

and

$$T^{-1} \sum Z_t.\Omega_{**}^{-1}.\kappa \begin{pmatrix} u_t \\ e_t \end{pmatrix} \xrightarrow{d} \left(\int W_e.\Omega_{**}^{-1}.\kappa.d\omega_\eta + Vec\left(\Delta_{e\eta}\kappa'\Omega_{**}^{-1}\right)\right) \tag{8.41}$$

where $W_e = I_k \otimes \omega_e$ where ω_e and ω_η are vectors of Brownian Motions. The construction of κ ensures that $\kappa\omega_\eta$ is independent of ω_e and therefore $\int W_e.\Omega_{**}^{-1}.\kappa.d\omega_\eta$ is mixed normal with mean zero and covariance matrix $\int W_e.\Omega_{**}^{-1} W_e'$. Therefore, given $N = T^{-1}I$ it follows that:

$$N \sum z_t \Omega_{**}^{-1} \kappa \begin{pmatrix} u_t \\ e_t \end{pmatrix} \xrightarrow{d} \mathcal{M}^{\mathcal{N}}\left(Vec\left(\Delta_{e\eta}\kappa'\Omega_{**}^{-1}\right), G^{*\prime}G^*\right). \tag{8.42}$$

It follows that v constructed as in [5.11] weakly converges to a multivariate normal.

REFERENCES

Andrews D.W.K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 817-858.

Apostol, T.M. (1974) Mathematical Analysis, Addison Wesley Publishing.

Balcombe .G. and Tiffin R.(2002). Fully Modified Estimation with Cross Equation Restrictions, *Economics Letters*, 74, 257-263

Berck P. and Sydsater K. (1993) Economists Mathematical Manual, Second Edition, Springer-Verlag.

Davidson, J. (1994) Stochastic Limit Theory, Advanced Texts in Econometrics, Oxford University Press.

Golan, A., Judge, G. and Miller, D. (1996). Maximum Entropy Econometrics, Robust Estimation with Limited Data. Series in Financial Economics and Quantitative Analysis. Wiley,

Golan, A. Judge, G. and Perloff, J. (1997). Estimation and inference with censored and ordered multinomial response data. *Journal of Econometrics* 79, 23-51.

Golan Amos, and Perloff J. (2002). Comparison of maximum entropy and higher-order entropy estimators, *Journal Of Econometrics* (107)1-2 195-211

Golan, A. Moretti E. and Perloff J.M. (1999). An Information Based Sample Selection Estimation of Agricultural Worker's Choice between Piece Rate and Hourly Work. *American Journal of Agricultural Economics*, Vol 81, 3, 735-741.

Golan, A. (2002). Information and Entropy Econometrics, Editors View. *Journal of Econometrics*, 107 1-15.

Golan A. and Gzyl H. (1999). A Generalized Maxentropic Inversion Procedure for Noisy Data. *Applied Mathematics and Computation* (forthcoming).

Hamilton J.D. (1994), Time Series Analysis, Princeton University Press, New Jersey.

Harmon, A. Preckel, P.V. and Eales J., (1998). Entropy Based Seemingly Unrelated Regression, Staff Paper #98-8, Dept of Agricultural Economics. Purdue University.

Haug A. (1999). Testing linear restrictions on cointegrating vectors: sizes and powers of Wald and Likelihood ratio tests in finite samples. Working Paper: University of Canterbury.

Kullback J. (1959). Information Theory and Statistics. John Wiley, New York.

Magnus J.R, and Neudecker H. (1994) Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley Series in Probability and Mathematical Statistics, Wiley and Sons..

Marsh L., R.C. Mittlehammer, and Cardell S. (1998). A Structural-Equation GME Estimator. A Selected Paper 1998 AAEA Annual Meeting, Salt Lake

Moon H.R. (1999). A note on fully-modified estimation of seemingly unrelated regressions models with integrated regressors. *Economics Letters 65*, 25-31.

Paris Q. (2001). MELE, Maximum Entropy Leuven Estimators, 2001, Working Paper, 01-003, California Agricultural Experiment Station, Gianinni Foundation for Agricultural Economics.

Prekel P.V. (2001). 'Least Squares and Entropy A Penalty Function Perspective', American Journal of Agricultural Economics, 83 (2) 366-377.

Phillips P.C.B. and Hansen B. 1990). Statistical inference in instrumental variable regressions with I(1) processes. *Review of Economics Studies 57*, 99-125.

Shannon C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379-423.

Thirtle, C., Sartorious von Bach, H. and van Zyl, J. (1993) Total Factor Productivity in South African Agriculture, 1947-1992. *Development South Africa*, 10, 301-318.

Xaio, Z and Phillips P.C.B. (2002) Higher order approximations for Wald Statistics in Time Series Regression with Integrated Processes. *Journal of Econometrics*, 108, 157-198.

Zellner, A. (1996) Models, prior information, and Bayesian Analysis. *Journal of Econometrics*, 75, 51-68.

Zellner A. (1997). A Bayesian Method of Moments (BMOM): Theory and Applications. Advances in Econometrics, vol 12. Applying Maximum Entropy to Econometric Problems: Eds T.B. Formby and R.C. Hill, pp. 85-105. Grenwich JAI Press, 1997.

Zellner, A. (1999) New Information Based Econometric Methods in Agricultural Economics: Discussion, *American Journal of Agricultural Economics*, Vol 81, 3, 742-46

Zellner A. (2002), Information Processing and Bayesian Analysis. *Journal of Econometrics*, 107, 41-50.