ISSN 1032-3813
ISBN 0 7326 0777 9

MONASH UNIVERSITY

AUSTRALIA

BAYESIAN ANALYSIS OF A COINTEGRATION MODEL USING MARKOV CHAIN MONTE CARLO

Gael Martin

Working Paper 16/95
October 1995

DEPARTMENT OF ECONOMETRICS

Bayesian Analysis of a Cointegration Model Using Markov Chain Monte Carlo*

Gael Martin
Department of Econometrics,
Monash University,
Clayton, Victoria 3168,
Australia.
Gael.Martin@BusEco.monash.edu.au

July 1995

Abstract

This paper presents a strategy for conducting Bayesian inference within the context of the triangular cointegration model. The numerical analysis is based on a hybrid of the Gibbs and Metropolis Markov Chain Monte Carlo methods. The use of a combination of two Markov Chain algorithms rather than a straight Gibbs Sampler occurs as a consequence of the complications induced by the prior specification. The specific form of the latter is, in turn, required for two purposes. First, in order to offset an identification problem which occurs when the cointegration model is extended to allow for the possibility of no cointegration. Second, in order to allow for an objective prior density on the parameter which determines the existence of cointegration.

---

*I would like to thank Brett Inder, Catherine Forbes, Herman van Dijk, Vance Martin, Brendan McCabe and Max King for helpful comments on an earlier draft of this paper. Any mistakes remain, of course, my own responsibility.

1 Introduction

In this paper, we propose a Bayesian approach to both testing for a cointegrating relationship between two or more time series and estimating that relationship, if it is deemed to exist. The method uses the triangular system representation of the cointegration model, as used extensively by Phillips and various co-authors in the Classical cointegration literature. (See, for example, Phillips 1991c and 1994 and Phillips and Hansen (1990)). The numerical technique used to produce estimates of the posterior densities of interest is a hybrid of the Gibbs and Metropolis Markov Chain Monte Carlo (MCMC) algorithms. (See Gelfand and Smith (1990) and Smith and Roberts (1993)).

We demonstrate that the use of the triangular system to model cointegration, in combination with the assumption of normally distributed errors, would potentially render flat prior inference on both the parameter controlling the presence of cointegration and the cointegrating parameter(s) themselves, extremely simple via a straight forward application of Gibbs Sampling. However, two complications arise.

First and most fundamentally, an identification problem will be shown to obtain when the triangular model of cointegration is extended to allow for a lack of cointegration, in order for an initial test of cointegration to be performed. Following an idea of Kleibergen and Van Dijk (1994a. and b.), we choose to solve this identification problem via the use of a Jeffreys' prior. This solution produces a joint posterior density, for which the associated conditional densities to be used in the Gibbs sampling are, in part, non-standard and, hence, difficult to simulate from. We choose to circumvent this problem by inserting, at the points in the Gibbs Sampler where simulation is difficult, sub-chains produced by an alternative Markov chain strategy, namely the Metropolis algorithm. We show that, as a consequence of the fact that the underlying identification problem does not impact at the level of the full conditional densities, there is an obvious choice of candidate density to be used in the Metropolis algorithm.

A second problem arises from an issue which has been discussed elsewhere in the literature (see, in particular, the Journal of Applied Econometrics, (1991)), namely the appropriate way in which to model a-priori non-informativeness in a time series context. As concerns a cointegration model, the point is that a flat prior on the parameter controlling the presence of cointegration is not a true representation of a-priori objectivity. Such a prior, in fact, serves to bias posterior inference in favour of cointegration. (See also Phillips (1993)). The incorporation of the appropriate objective prior introduces a non-standard aspect into the Gibbs strategy. The management of this induced complication is also possible via the imbedding of a Metropolis sub-chain, although the choice of candidate density may become more problematic.

The paper is organized as follows. Section 2 provides an outline of both the model and the inferential objectives. Section 3 demonstrates the identification problem which arises and its solution via the Jeffreys' prior principle. The proposed hybrid Gibbs/Metropolis sampling method is then described in Section 4, along with an informal discussion of the required convergence criteria. In Section 5, we demonstrate the ability of the proposed MCMC method to reproduce the exact marginal densities. We also provide the results of a small Monte Carlo study, in which the Bayesian inferences are compared with Classical alternatives. Although preliminary, these results tend to suggest that the Bayesian method provides a very viable alternative to the Classical procedures. The paper then gives some conclusions in Section 6. An outline of the formal conditions for the convergence of the Gibbs and Metropolis Markov chain algorithms is provided in the Appendix.

2 The Model and the Inferential Objectives

Consider the following bivariate model for the generation, at time t, of observations on the variables y and x respectively:

$$y_t = \beta x_t + u_{1t} \tag{1}$$

$$x_t = x_{t-1} - u_{2t} \tag{2}$$

The error vector $u_t = (u_{1t}, u_{2t})'$ is assumed to be an autoregressive process, characterized as $B(L)u_t = e_t$, where $B(L)$ is a matrix of finite degree, one-sided polynomials in the lag operator L of the form:

$$B(L) = \begin{bmatrix} b_{11}(L) & b_{12}(L) \\ b_{21}(L) & b_{22}(L) \end{bmatrix}, \tag{3}$$

where:

$$b_{ij}(L) = b_{ij0} - \sum_{k=1}^{p_{ij}} b_{ijk} L^k, \qquad i,j = 1,2,$$

with $b_{ij0} = 1$ unless all coefficients are set to zero, in which case $b_{ij0} = 0$ also. It is assumed that $e_t = (e_{1t}, e_{2t})'$ has a bivariate Normal distribution of the form:

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} \sim NID(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}), \qquad \sigma_{12} = \sigma_{21}. \tag{4}$$

Given the $I(1)$ nature of the regressor, (1) potentially represents a cointegrating relationship between two $I(1)$ variables. It is of interest both to test for this possibility and to estimate the value of β in the event that the latter is concluded to be the parameter of a cointegrating relationship.

There are several points to make regarding the chosen model specification. The first concerns the off-diagonal terms in the $B(L)$ matrix. It can be shown that the presence of a non-zero $b_{21}(L)$ polynomial has non-trivial implications for the presence or not of a cointegrating relationship between x_t and y_t . In addition, the allowance of a non-zero $b_{12}(L)$ polynomial, produces an identification problem which does not seem to be solvable via the method to be outlined in the paper. As such, we begin by imposing the restriction of a diagonal $B(L)$ matrix.

The specification of a diagonal $B(L)$ matrix obviously imposes a simpler autocorrelation structure on u_{1t} and u_{2t} . In addition, the restriction $b_{21}(L) = 0$, in particular, implies that x_t is *strongly exogenous.*[1] We believe, however, that the allowance of a certain level of endogeneity of x_t via a non-diagonal Σ matrix. plus the allowance of general AR specifications for $b_{11}(L)$ and $b_{22}(L)$ provides a rich enough parameterization for the model to be useful. Methods for handling the model with a full $B(L)$ matrix are the subject of current research.

The basic mechanics of the procedure demonstrated in the paper remain the same when the model is altered in such a way that (1) represents either a multiple regression equation or a multivariate system of equations (in which case y_t becomes a vector and B a matrix), so long as the particular structure of the model is maintained. As a consequence, we are justified in simplifying the exposition by presenting results for the bivariate case.

---

[1] See Engle, Hendry and Richard (1983).

The fundamentals of the method are invariant to the inclusion of a drift and/or a deterministic trend term in the equation for x_t . As a consequence, we abstract from these elements. The inclusion of a constant term in the model does, however, have a significant impact on the procedure. Whilst the additional identification problem which it induces can be tackled via a straight forward extension of the Jeffreys' prior used to solve the fundamental identification problem in the model, the impact of the lack of identification at the level of the full conditional densities alters the convergence properties of the proposed Gibbs - based MCMC scheme. In order to limit the scope of the paper, we do not consider this problem in any detail.

2.1 Inference regarding the presence of cointegration

The first step in the inferential process is to determine whether the error term, u_{1t} , is stationary and the associated regression equation for y_t and x_t a cointegrating one as a consequence. Given that an assessment of cointegration is to be performed, the prior density for the parameters determining the autocorrelation structure of u_{1t} will allow for non-stationarity in the latter. The assessment will then be based upon the relative probabilities of stationarity and non-stationarity as calculated from the marginal density of the parameter controlling the unit root. This approach thus requires a reparameterization of the autocorrelation structure in u_{1t} along the lines used in an augmented Dickey Fuller test, so that the single marginal density associated with the parameter controlling the unit root is the basis of inference.

The equating of cointegration and non-cointegration respectively with the relevant marginal probabilities of stationarity and non-stationarity, does render the nature of the testing problem slightly different from that of a standard Classical residual-based cointegration test. If the probability of stationarity is high enough for the hypothesis of cointegration to be "accepted", then one is indeed in the standard situation, whereby two $I(1)$ variables have been found to possess an $I(0)$ linear combination. If, on the other hand, the error term is deemed to be non-stationary, then the regression structure of (1) implies that y_t is explosive.

Two points need to be made here. First, in order for such a finding to be consistent with the pre-testing on y_t and x_t which has, by assumption, preceded the cointegration analysis, it would appear to be necessary to use a Bayesian unit root test in which the possibility of an explosive y_t (and x_t) is allowed for. Second, in order to preclude a-priori the possibility of a highly explosive error term and, hence, a y_t series which would not be a contender in terms of cointegration with x_t , the marginal prior on the unit root parameter in the error term may well need to be truncated at some point beyond 1. This latter point shall be raised again in Section 3.

2.2 Estimation of the cointegrating parameter

Once the decision has been made as to the likelihood of a cointegrating relationship existing, inference regarding the parameter of that relationship may proceed, based on the marginal density function for β. The prior on the parameter controlling the stationarity in u_{1t} could be restricted to the stationary part of the parameter space, in accordance with the finding of cointegration, if so desired. We choose to demonstrate the proposed inferential method without the restriction of cointegration imposed at any point, in order to simplify the exposition. Only a minor modification of the numerical technique used would be needed to incorporate the cointegration restriction.

3 An Identification Problem and its Solution via the Jeffreys' Prior Principle.

In their 1994 papers, Kleibergen and Van Dijk encounter an identification problem, the consequence of which is the lack of proper marginal posterior densities for use in Bayesian inference. The solution which they propose involves the application of the Jeffreys' principle for prior density elicitation.[2]

An identification problem also arises within the context of the present model specification, with the Jeffreys' prior principle again providing the remedy. The nature of the problem is most easily demonstrated within the context of a very reduced form of the model with $b_{11}(L) = 1 - \phi_1 L$ and $b_{22}(L) = 1$.[3] The features of the problem and the solution to it that we suggest remain fundamentally the same when u_{1t} and u_{2t} are allowed to be general AR processes. The restrictions $b_{12}(L) = b_{21}(L) = 0$ will be imposed from this point onwards.

We shall assume prior independence between the parameter matrix Σ and the remaining parameters in the model, namely β and ϕ_1. As such, we can

---

[2] See also Schotman and Van Dijk (1991).

[3] The notation ϕ_1 is used in order to be consistent with that used when $b_{11}(L)$ is reparameterized later in the Section.

decompose the joint prior for all parameters as:

$$p(\Sigma,\beta,\phi_1) \propto p(\Sigma).p(\beta,\phi_1) \ . \tag{5}$$

As we shall see, it is the form assumed by the second component in (5) which has implications for parameter identification. For the time being, we shall allow this component to be a uniform, or flat, density function. For Σ we shall utilize the noninformative Jeffreys' prior, $|I_\Sigma|^{1/2} \propto |\Sigma|^{-3/2}$, where I_Σ denotes the submatrix of the information matrix which relates to the elements of Σ; i.e. $I_\Sigma = E(-\partial^2 \ln L/\partial\Sigma\partial\Sigma')$. This particular prior on Σ, in combination with the joint flat prior on β and ϕ_1, allows standard Bayesian analysis to be performed, with the precise nature of the identification problem able to be easily highlighted.

Given the distributional assumptions underlying the model and the choice of priors as discussed above, the form of the joint posterior density is:

$$\begin{aligned} p(\beta,\phi_1,\Sigma) &\propto |\Sigma|^{-(n+3)/2} . \exp\{-1/2(\sum_t e_t\Sigma^{-1}e_t)\} \\ &\propto |\Sigma|^{-(n+3)/2} . \exp\{-1/2tr(\Sigma^{-1}S)\}, \end{aligned} \tag{6}$$

with (β,ϕ_1,Σ) defined on $D = \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{S}^{pds}$, where $\mathbb{S}^{pds}$ denotes the space of (2×2) positive definite symmetric matrices. The vectors $\mathbf{x}$ and $\mathbf{y}$ denote the n-dimensional observation vectors and $S = \sum_t e_t e_t'$. In the simplified version of the model, $e_{1t} = (1-\phi_1 L)u_{1t} = (1-\phi_1 L)(y_t - \beta x_t)$ and $e_{2t} = u_{2t}$.

Standard techniques enable integration with respect to Σ, yielding the following form for the joint density of β and ϕ_1:

$$p(\beta,\phi_1|\mathbf{y},\mathbf{x}) \propto \{C_1 + C_2\beta^2 - 2C_3\beta\}^{-n/2}, \tag{7}$$

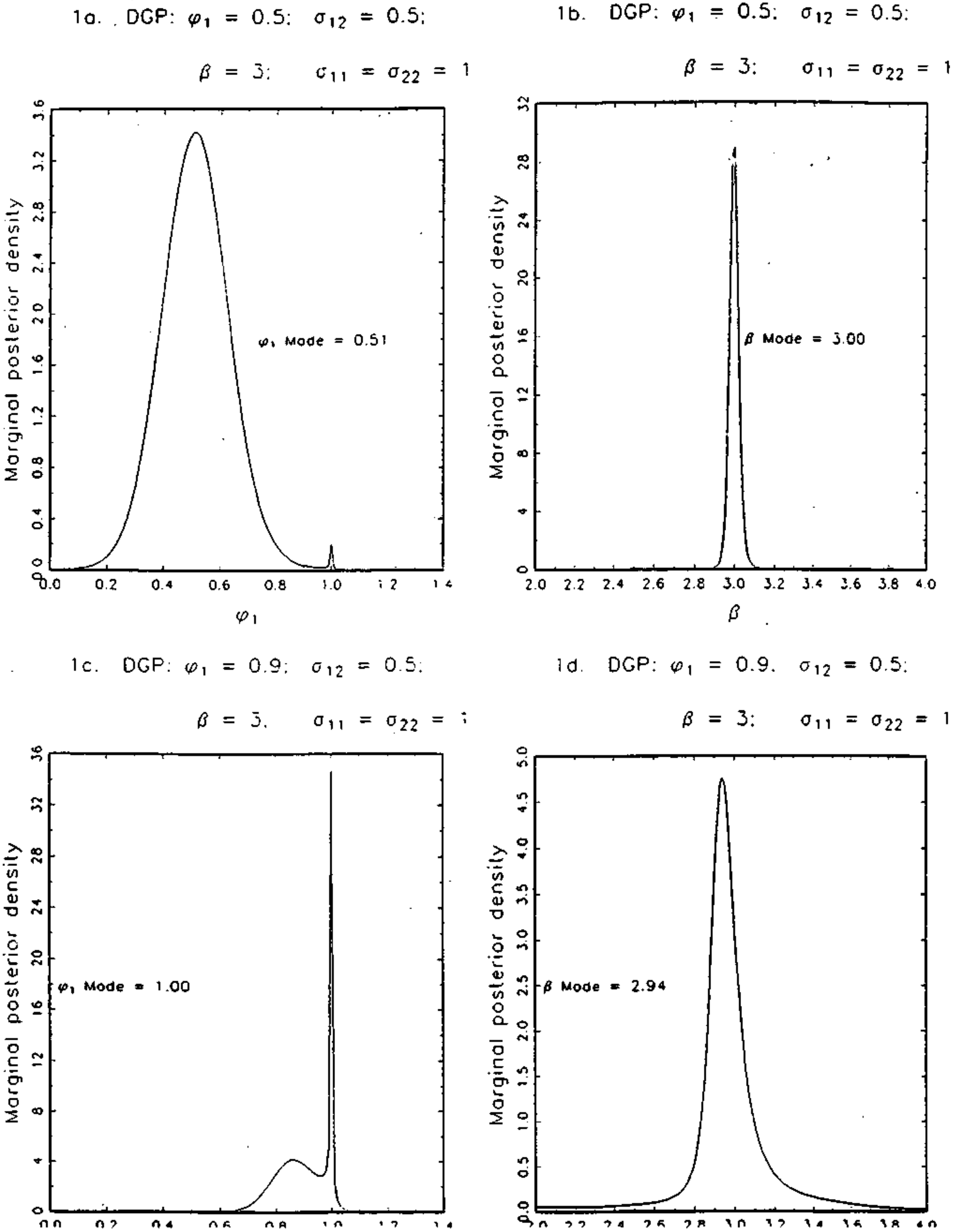
where:

$$\begin{aligned} C_1 &= \sum_t[(1-\phi_1 L)y_t]^2 \sum_t \Delta x_t^2 - [\sum_t(1-\phi_1 L)y_t\Delta x_t]^2, \\ C_2 &= \sum_t[(1-\phi_1 L)x_t]^2 \sum_t \Delta x_t^2 - [\sum_t(1-\phi_1 L)x_t\Delta x_t]^2, \\ C_3 &= \sum_t(1-\phi_1 L)y_t(1-\phi_1 L)x_t \sum_t \Delta x_t^2 - \\ &\quad [\sum_t(1-\phi_1 L)y_t\Delta x_t].[\sum_t(1-\phi_1 L)x_t\Delta x_t] \end{aligned}$$

and we further define:

$$C_4 = C_1 - C_3^2/C_2.$$

Figure 1. Marginal posterior densities for ϕ_1 and β based on a joint flat prior for β and ϕ_1.



1a. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$

1b. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$

1c. DGP: $\varphi_1 = 0.9$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$

1d. DGP: $\varphi_1 = 0.9$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$

It can be seen that when $\phi_1 = 1$, both C_2 and $C_3 = 0$. This in turn implies that the joint density for β and ϕ_1 is a constant function of and, hence, fails to identify β, when $\phi_1 = 1$. The integral, with respect to β, of this slice of the joint density, being unbounded. ascribes an infinite value to the marginal density of ϕ_1 at the point $\phi_1 = 1$.

For $\phi_1 \neq 1$, standard integration techniques can again be used to produce respectively the conditional and marginal densities:

$$p(\beta|\phi_1, \mathbf{y}, \mathbf{x}) \propto [\mathrm{s}_\beta^2]^{-1/2}\{1 + (\beta - \bar{\beta})^2/[(n-1)s_\beta^2]\}^{-n/2} \tag{8}$$

and:

$$p(\phi_1|\mathbf{y}, \mathbf{x}) \propto C_2^{-1/2} C_4^{-(n-1)/2} \tag{9}$$

where $\bar{\beta} = C_3/C_2$ and $(n-1)s_\beta^2 = C_4/C_2$. L'Hopital's rule can be applied to the relevant quantities in (8) and (9) in order to ascertain the nature of these densities as ϕ_1 approaches 1. We find that the Student t density for β conditional on ϕ_1 has both mean and variance approaching ∞ as $\phi_1 \rightarrow 1$. The marginal density of ϕ_1 approaches ∞ as $\phi_1 \rightarrow 1$.

If the impact of the lack of identification of β at $\phi_1 = 1$ were to be felt only at that single point, then the solution would be to simply redefine the joint density function as (6) with support $D^* = D \cap \{(\beta, \phi_1, \Sigma); \phi_1 \neq 1\}$. This amounts to finding a joint density which is equivalent almost everywhere (with respect to Lebesque measure) to the original density, but which does not incorporate the identification problem. Our inferences regarding both β and ϕ_1 would be unaffected by such a change in the definition of the joint posterior.

What we find, however, is that the impact of a "near" lack of identification can be significant for a wide range of ϕ_1 values around 1, depending on the nature of the true underlying data generating process (dgp). Figures 1a and 1c provide examples of the marginal density of ϕ_1 when the data has been generated from processes with true values of 0.5 and 0.9 respectively for ϕ_1. As is quite evident, when the bulk of the density is situated well below $\phi_1 = 1$, as in Figure 1a, there is only a slight asymptoting effect on the density for values of ϕ_1 around 1. In the case of Figure 1c. however, the distortion of the density is significant, with an artificial global mode being produced at a point arbitrarily close to 1.[4] Figures 1b. and 1d. present the corresponding marginal β densities, the flat tails of the latter density indicating the potential impact of the near identification problem on the existence of marginal moments for β.

We can shed more light on the impact of this near lack of identification on the marginal density of ϕ_1 specifically, by analyzing more closely the two quantities, $C_2^{-1/2}$ and $C_4^{-(n-1)/2}$, of which $p(\phi_1|\mathbf{y},\mathbf{x})$ is comprised. It is straight forward to show that C_2 is a quadratic function of ϕ_1 which assumes a minimum value of zero at $\phi_1 = 1$, irrespective of the true value of ϕ_1 in the underlying dgp. The only data dependent aspect of C_2 (and, hence $C_2^{-1/2}$) is the degree of concentration of the function around its minimum (maximum) value, and that is, in turn, affected by $\mathbf{x}$ only.

C_4 on the other hand, is a quadratic function of ϕ_1 whose minimum value occurs at a value of related to the OLS estimator of ϕ_1 in the hypothetical regression model $u_{1t} = u_{1t-1} + e_{1t}$ and which, as a consequence, is directly affected by the true value of ϕ_1.

In summary, the marginal density of ϕ_1 is the product of a well-behaved function of ϕ_1, $C_4^{-(n-1)/2}$, which possesses sensible inferential content regarding ϕ_1, and a function, $C_2^{-1/2}$, which possesses no such content. The latter, moreover, serves to dominate the former function, for certain dgp's, producing a density with a large amount of probability content around $\phi_1 = 1$, even when the true ϕ_1 is well into the stationary region. It would appear to be desirable, therefore, to somehow offset the $C_2^{-1/2}$ factor, in order to produce sensible inferences.

As has been alluded to several times now, the elimination of the impact of the identification problem, of which the distortion induced by $C_2^{-1/2}$ is the manifestation, may be achieved via the use of a particular Jeffreys' prior. The latter is proportional to the determinant of the information matrix and, hence, related to the inverse of the covariance matrix of the relevant posterior density. In the case of potentially unidentified parameters, it should tend to offset the infinite conditional variances which occur at the points at which the parameters become unidentified, as well as eliminating any associated irregularity in the marginal densities in regions of a near lack of identification.

---

[4] These densities are typical of densities resulting from the specified types of dpg's. Numerical integration was used to produce these densities and those in Figure 2. The MCMC algorithm is not applied until Section 4. The value $\phi_1 = 1$ has been eliminated from the support of the density for the purpose of producing the graphs.

As pointed out by Kleibergen and Van Dijk, the success of the Jeffreys' prior in this regard depends crucially on the way in which the expectations within it are evaluated.

In order to derive the appropriate form of the Jeffreys' prior, it is necessary to allow the identification problem to reveal itself in the full likelihood function. That is, the relevant parameter must fail to be identified by the full likelihood function in order for the Jeffreys' prior, as derived from that function, to be operational in terms of offsetting the lack of identification. In our case, the identification problem is revealed only after integration with respect to Σ. However, a simple decomposition of the full likelihood along the following lines enables the lack of identification of β when $\phi_1 = 1$ to manifest itself appropriately:

$$\begin{aligned} L(\Sigma, \beta, \phi_1 | \mathbf{y}, \mathbf{x}) &\propto |\Sigma|^{-n/2} . \exp\{-1/2tr(\Sigma^{-1}S)\} \\ &\propto \sigma_{11.2}^{-n/2} . \exp\{-1/(2\sigma_{11.2}) \sum_t [(y_t^* - \beta x_t^*) - (\sigma_{12}/\sigma_{22})\Delta x_t]^2\} \\ &\qquad . \sigma_{22}^{-n/2} \exp\{-1/(2\sigma_{22}) \sum_t [\Delta x_t]^2\}, \end{aligned} \tag{10}$$

where $y_t^* = y_t - \phi_1 y_{t-1}$, $x_t^* = x_t - \phi_1 x_{t-1}$ and $\sigma_{11.2} = \sigma_{11} - \sigma_{12}^2/\sigma_{22}$. Maintaining the assumption of prior independence of Σ and the remaining parameters, the first line of (10) can be used to derive the Jeffreys prior for Σ, namely, $|\Sigma|^{-3/2}$. With this independence assumed, the element σ_{12}/σ_{22} appearing in the second line of (10) can be replaced by the artificial parameter α, and the Jeffreys prior for β and α conditional on ϕ_1 can be derived from this first part of the decomposition, being the only part of (10) in which these parameters appear.

We wish therefore to derive the determinant of the (2× 2) information matrix:

$$I_{\beta,\alpha} = E \begin{bmatrix} -\partial^2 \ln L/\partial\beta^2 & -\partial \ln L/\partial\beta\partial\alpha \\ -\partial^2 \ln L/\partial\alpha\partial\beta & -\partial^2 \ln L/\partial\alpha^2 \end{bmatrix}, \qquad -\partial^2 \ln L/\partial\beta\partial\alpha = \partial^2 \ln L/\partial\alpha\partial\beta.$$

where it is implicit that all differentiation is conditional on ϕ_1. It is simple to show that the elements of this matrix reduce to:

$$\begin{aligned}
E(-\partial^2 \ln L/\partial\beta^2) &= (1/\sigma_{11.2})E\sum_t (x_t - \phi_1 x_{t-1})^2 &&= (1/\sigma_{11.2})E(\mathbf{x}^{*\prime}\mathbf{x}^*),\\
E(-\partial^2 \ln L/\partial\beta\partial\alpha) &= (1/\sigma_{11.2})E\sum_t (x_t - \phi_1 x_{t-1})\Delta x_t &&= (1/\sigma_{11.2})E(\mathbf{x}^{*\prime}\Delta\mathbf{x})\\
&\textit{and}:\\
E(-\partial^2 \ln L/\partial\alpha^2) &= (1/\sigma_{11.2})E\sum_t \Delta x_t^2 &&= (1/\sigma_{11.2})E(\Delta\mathbf{x}'\Delta\mathbf{x}),
\end{aligned}$$

where $\mathbf{x}^*$ and $\Delta\mathbf{x}$ are the observation vectors for x_t^* and $\Delta x_t = x_t - x_{t-1}$ respectively. As such, the Jeffreys' prior, which is proportional to the square root of the determinant, is defined by:

$$|I_{\beta,\alpha}|^{1/2} \propto \{E(\mathbf{x}^{*\prime}\mathbf{x}^*).E(\Delta\mathbf{x}'\Delta\mathbf{x}) - [\mathbf{E}(\mathbf{x}^{*\prime}\Delta\mathbf{x})]^2\}^{1/2}.$$

A marginal Jeffreys' prior on ϕ_1 can also be defined as $I_{\phi_1}^{1/2} \propto \{E(-\partial \ln L/\partial\phi_1^2)\}^{1/2}$. With the first exponent term in (10) once again being the only relevant term, the form of the latter prior can be derived as $\{E\sum_t(u_{1t-1}^2)\}^{1/2} = \{E(\mathbf{u}_{1,-1}'\mathbf{u}_{1,-1})\}^{1/2}$, with $\mathbf{u}_{1,-1} = (u_{10}, u_{11}, \ldots, u_{1n-1})'$.

The combination of the conditional and marginal priors then leads to an implied joint prior for α, β, and ϕ_1 and of the form:

$$\{E(\mathbf{x}^{*\prime}\mathbf{x}^*).E(\Delta\mathbf{x}'\Delta\mathbf{x}) - [\mathbf{E}(\mathbf{x}^{*\prime}\Delta\mathbf{x})]^2\}^{1/2}.\{E(\mathbf{u}_{1,-1}'\mathbf{u}_{1,-1})\}^{1/2} \qquad (11)$$

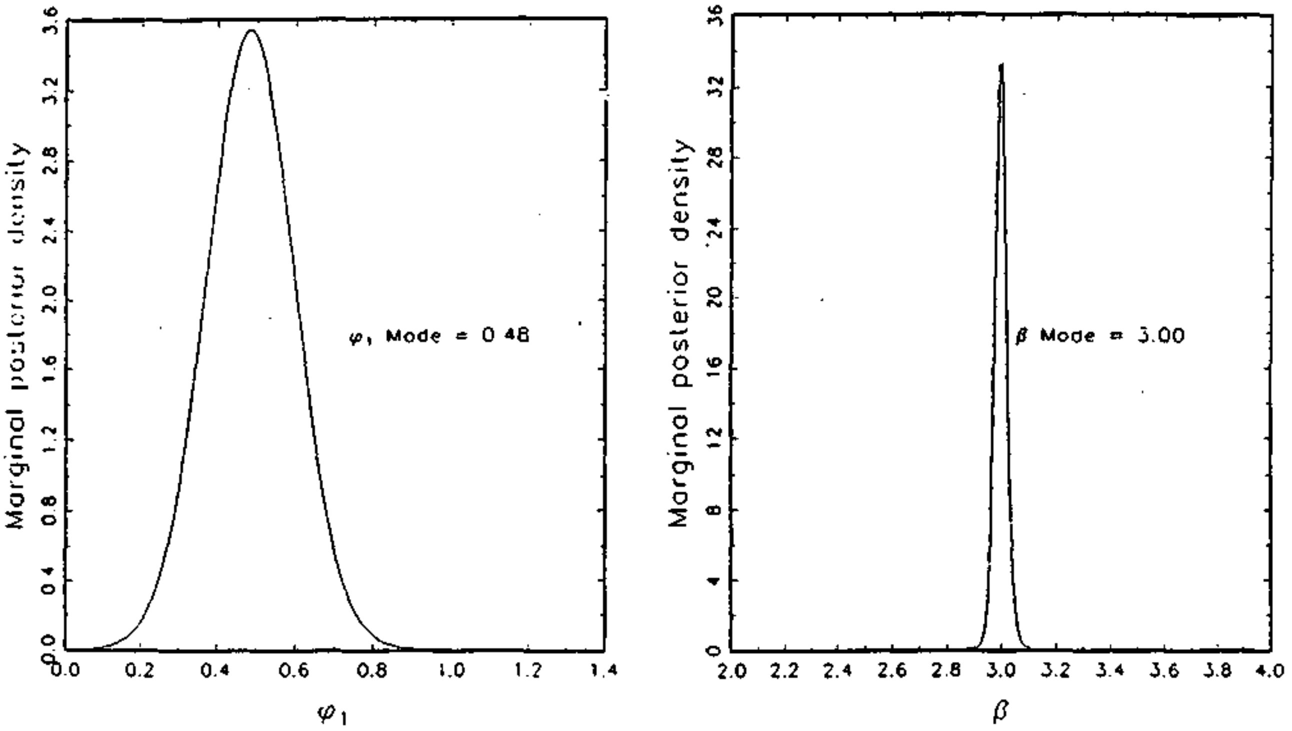
Given that α is only an artificial parameter introduced for the purposes of the derivation, we can view the above prior in terms of the decomposition: $p(\beta|\phi_1).p(\phi_1)$. The implied marginal prior on β is obviously flat.

The source of the distortion in the ϕ_1 density, namely the quantity $C_2^{-1/2}$ is equivalent to the inverse of the first bracketed part of (11), *so long as the expectations of the functions of x_t which appear are replaced by their realized values.* Given that x_t is weakly exogenous with respect to both β and ϕ_1, by virtue of the imposition of a zero value for $b_{21}(L)$, this form of evaluation of the expectations implies no loss of information with regard to the parameters of interest. This particular version of the Jeffreys' prior is then sufficient to exactly offset the impact of the identification problem. Given that we shall also be allowing for the marginal Jeffreys' prior on ϕ_1, we shall sometimes refer to this conditional Jeffreys' prior as the "smoothing" prior, so as to avoid possible confusion.

Figure 2. Marginal posterior densities for ϕ_1 and β based on the conditional Jeffreys' prior for β given ϕ_1.

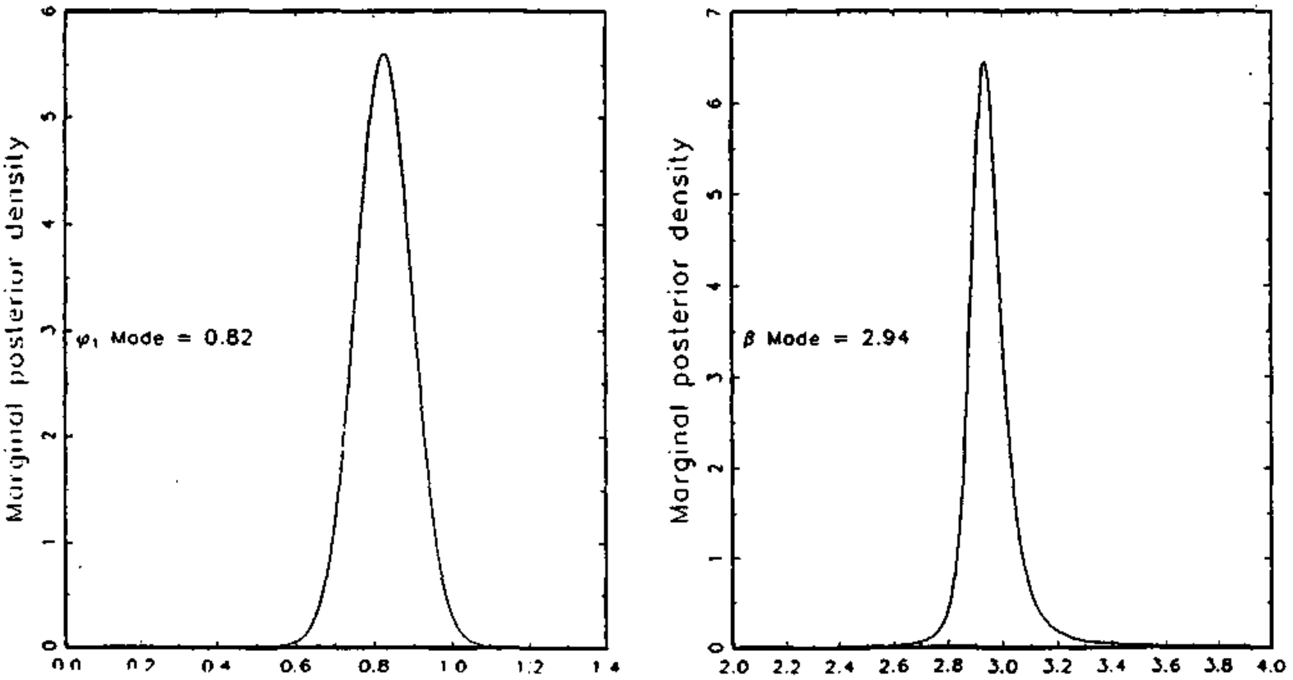
2a. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$

2b. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$



26. DGP: $\varphi_1 = 0.9$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$

2b. DGP: $\varphi_1 = 0.9$; $\sigma_{12} = 0.5$; $\beta = 3$; $\sigma_{11} = \sigma_{22} = 1$



Figures 2a and 2c present the "smoothed" versions of the densities in Figures 1a and 1c respectively.[5] As would be anticipated, given the small impact of the $C_2^{-1/2}$ factor on the ϕ_1 density given in Figure 1a, the eradication of it in Figure 2a affects the density only slightly. The same can be said of the corresponding β density in Figure 2b. The difference between the ϕ_1 densities in Figures 1c and 2c, however, is much more marked, with the latter displaying nothing of the irregularity present in the former. The impact on the β density, as seen in Figure 2d, is to produce better tail behaviour. The β mode, it will be noted, is not affected by the smoothing process.

As to the expectation appearing in the marginal prior for ϕ_1, its evaluation as a realized value has more significant implications for the nature of the prior information being specified about the parameters. If the expectation of $\mathbf{u}'_{1,-1}\mathbf{u}_{1,-1}$ is taken to be equivalent to the sample value, then the temporal dependence of u_t on ϕ_1 is being ignored and the resultant prior on ϕ_1 is flat. If, on the other hand, the time series structure of u_{1t} is explicitly taken into account, then the form of the marginal prior for ϕ_1 (given the assumption of a value of 0 for u_0) is given by:

$$\begin{aligned} p(\phi_1) &\propto \{E[\sum_t u_{1t-1}^2]\}^{1/2} \\ &\propto \begin{cases} \{[1/(1-\phi_1^2)][n-(1-\phi_1^{2n})/(1-\phi_1^2)]\}^{1/2} & \phi_1 \neq 1 \\ \{n(n-1)/2\}^{1/2} & \phi_1 = 1 \end{cases} \end{aligned} \tag{12}$$

The form of this density (for n = 50) is given in Figure 3.

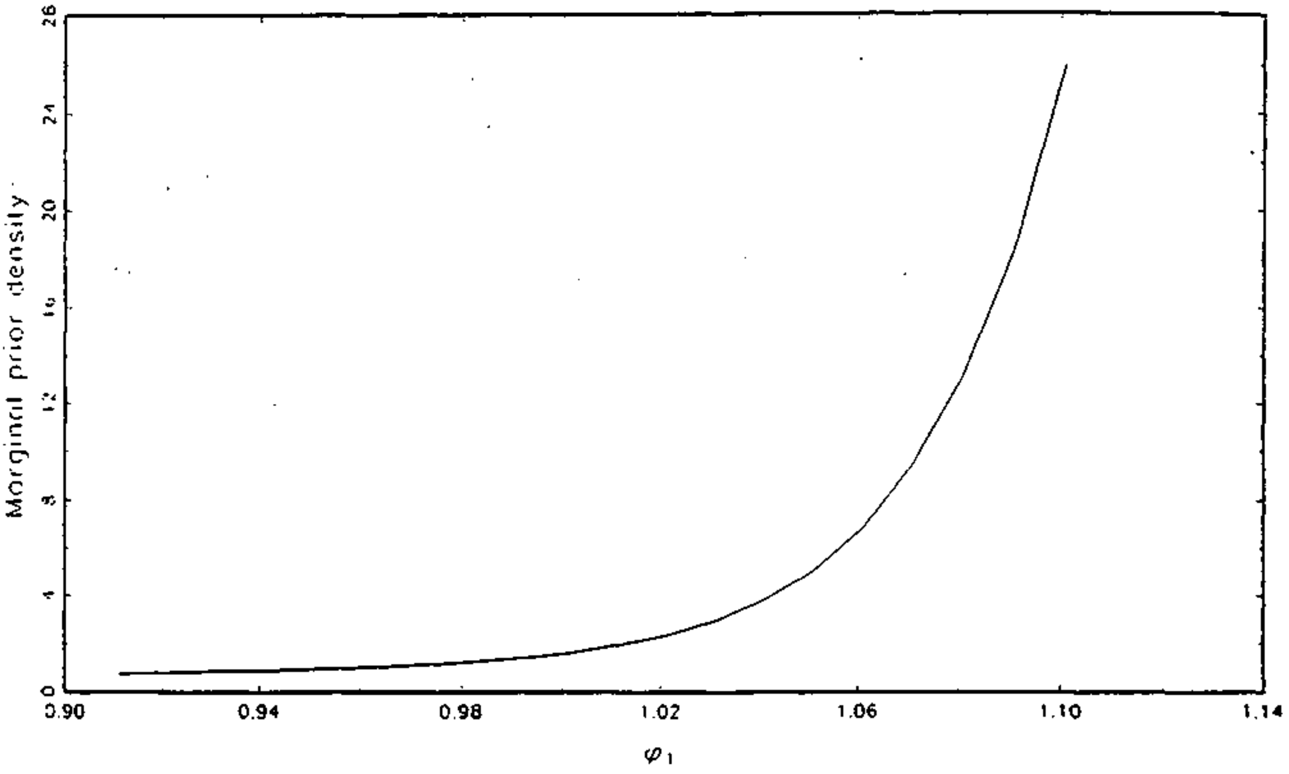
As discussed by Phillips (1991a and b), (12) is the true Jeffreys' prior and, as such, is noninformative or objective in the sense that such a prior is.[6] One manifestation of this noninformativeness is the way in which the Jeffreys' prior, as depicted in Figure 3, being a function as it is of the structure of the underlying model, reflects that model in an appropriate manner. In particular, the high weight given to large values of ϕ_1 by the Jeffreys' prior reflects the fact that with an AR(1) model, the data would provide much more information about ϕ_1 as ϕ_1 increased.

---

[5] The marginal prior on ϕ_1 is still flat at this stage.

[6] See Box and Tiao (1973, Chp1) for one interpretation of the sense in which the Jeffreys' prior is "noninformative".

Figure 3. Marginal Jeffreys' prior for ϕ_1.



Given the shape of the marginal Jeffeys' prior. it is obvious that its use would produce a marginal posterior density for ϕ_1 with more probability content in the non-stationary region than would the use of a flat prior on ϕ_1. Given that ϕ_1 is the parameter of an AR(1) error process, one could argue a large mass of probability associated with very large values of ϕ_1 is not appropriate. That is, a-priori, we are likely to believe that a highly explosive error process is an unlikely feature of a sensibly specified cointegrating regression.

Phillips (1993) suggests a modification of the Jeffreys' prior which reflects this a-priori reasoning by assigning low probability to high values of ϕ_1. Some thought, however, needs to be given to the extent of "truncation" required in any particular instance. As such, once again in order to limit the scope of the paper, we compare inferences based only on the flat and unrestricted Jeffreys' priors for ϕ_1.

The prior appropriate for the model incorporating general AR polynomials $b_{11}(L)$ and $b_{22}(L)$, of orders p_{11} and p_{22} respectively, is derived in precisely the same way as (11). In order to characterize a unit root in u_{1t} in terms of the single parameter ϕ_1, $b_{11}(L)u_t$ and $b_{11}(L)x_t$ are reparameterized respectively as:

$$\begin{aligned} b_{11}(L)u_{1t} &= u_{1t} - \phi_1 u_{1t-1} - \phi_2(u_{1t-1} - u_{1t-2}) - \cdots - \phi_{p_{11}}(u_{1t-p_{11}+1} - u_{1t-p_{11}}) \\ &= u_{1t} - \phi' \mathbf{u}^r_{1t-1} \\ &= u^*_{1t} \end{aligned}$$

and:

$$\begin{aligned} b_{11}(L)x_t &= x_t - \phi_1 x_{t-1} - \phi_2(x_{t-1} - x_{t-2}) - \cdots - \phi_{p_{11}}(x_{t-p_{11}+1} - x_{t-p_{11}}) \\ &= x_t - \phi' \mathbf{x}^r_{t-1} \\ &= x^*_t \end{aligned}$$

where:

$$\phi_1 = b_{111} + b_{112} + \cdots + b_{11p_{11}}$$

is a measure the long-term memory of the process, and the remaining parameters are defined by $\phi_k = -\sum_{k=2}^{p_{11}} b_{11k}$, $k = 2, 3, \ldots, p_{11}$. In addition, we express $b_{22}(L)\Delta x_t$ as:

$$\begin{aligned} b_{22}(L)x_t &= x_t - x_{t-1} - b_{221}(x_{t-1} - x_{t-2}) - \cdots - b_{22p_{22}}(x_{t-p_{22}+1} - x_{t-p_{22}}) \\ &= x_t - (1, \mathbf{b}'_{22})\mathbf{x}^r_{t-1} \\ &= x^{**}_t. \end{aligned}$$

Using this parameterization (and the obvious associated vector notation), the resultant prior specification is of the form:

$$\{(\mathbf{x}^{*\prime}\mathbf{x}^*).(\mathbf{x}^{**\prime}\mathbf{x}^{**}) - (\mathbf{x}^{*\prime}\mathbf{x}^{**})^2\}^{1/2}.\{E\sum_t \left|\mathbf{u}_{1t-1}\mathbf{u}'_{1t-1}\right|\}^{1/2} \tag{13}$$

The first bracketed term is derived in exactly the same fashion as the corresponding term in (11), except that the conditioning is now on ϕ and b_{22}. Once again, the expectations with respect to x_t are equated with the realized values, in order to ensure that this factor exactly offsets the identification problem which occurs in the more general specification when the two terms x^*_t and x^{**}_t coincide. This coincidence of values occurs at the points in the joint parameter space where $\phi_1 = 1$, $\phi_i = b_{22i-1}$, $i = 2, 3, \ldots, \min(p_{11}, p_{22})$ and any excess elements in ϕ or $\mathbf{b}_{22}$ equal zero.

The second term in (13) is an obvious extension of the corresponding term in (11), whilst the implicit third term, namely the marginal prior for $\mathbf{b}_{22}$ is flat. The reason for the latter is that the marginal Jeffreys' prior for $\mathbf{b}_{22}$ is a function of expectations of x_t only. With these expectations being required to be set to the realized values, the prior becomes a constant function with respect to $\mathbf{b}_{22}$. The use of a flat prior for this set of time series parameters is obviously in conflict with the allowance of a non-flat prior for ϕ.

Before moving on to consideration of the details of the numerical method to be applied in the paper, it is of interest to compare our inferential approach with that of one particular Classical single equation method, namely that of Phillips and Loretan (1991). Couching their method within the context of the simple model initially considered in this Section, it involves the application of non-linear least squares (NLS) to the transformed and augmented model:

$$y_t^* = \beta x_t^* + \alpha \Delta x_t + \eta_t \tag{14}$$

Conditional on ϕ_1, the OLS estimator of β is equivalent to the mean of our conditional density of β given ϕ_1. When $\phi_1 = 1$, (14) exhibits perfect multicollinearity in the regressors and the OLS estimator of β is undefined. This is simply a different manifestation of the identification problem encountered in our approach, with near perfect multicollinearity likely to have a similar impact on the Classical estimates as a near lack of identification does on the flat prior Bayesian inferences. That is, although Classical inference proceeds on the assumption that β is the parameter of a cointegrating relationship and that ϕ_1 is less than 1, so long as the criterion function to be optimized does not incorporate this restriction, then it will be flat in the directions of β and α when $\phi_1 = 1$. Even if this restriction were imposed, it would not eliminate the problem of the likelihood being close to flat in regions of near multicollinearity. If this part of the parameter space were empirically significant, then the effect would be to produce convergence problems and inefficient estimates. Since we explicitly adjust for the identification problem in our method, we would anticipate that our inferences would be more accurate than those produced by the Classical procedure.

Another way of looking at this is to compare the mean of the marginal density of β, $E(\beta|\mathbf{y}, \mathbf{x}) = \int E(\beta|\phi_1, \mathbf{y}, \mathbf{x}) p(\phi_1|\mathbf{y}, \mathbf{x}) d\phi_1$, as an estimator of β. with the estimator of Phillips and Loretan. Whilst the Classical estimator is equivalent to the conditional mean above, conditional on one particular estimate of ϕ_1 as produced in the estimation process, the marginal mean is the average of the conditional mean with respect to the marginal density of ϕ_1. It is this marginal posterior of ϕ_1 which contains in it the factor $C_2^{1/2}$ which adjusts for the identification problem. As such, we would expect the marginal mean and, in the case of a symmetric density, any other location measure, to be a more efficient estimator then the Classical alternative. The Monte Carlo results which we present in Section 5 certainly confirm this intuition.

4 The Theory and Mechanics of the Gibbs Sampling and Metropolis algorithms

The proposal of this paper is to produce estimates of the marginal densities of interest, namely those of β and the unit root parameter via a combination of MCMC sampling strategies. A brief explanation of the general nature of the two sampling schemes to be used, Gibbs and Metropolis, shall be given, followed by a more detailed description of their application to the particular model at hand. With reference made to the formal convergence theory presented in the Appendix, we then demonstrate that the hybrid scheme does appear to satisfy the various conditions for convergence to the joint posterior distribution. For recent papers discussing both the theory and implementation of MCMC procedures see Tierney (1991), Smith and Roberts (1993) and Roberts and Smith (1994). The book by Tanner (1994) also provides informative and comprehensive discussion of the methods within the broader context of Bayesian computational methods.

4.1 The Gibbs Sampling Algorithm

As applied in a Bayesian context, Gibbs sampling involves an iterative generation of random drawings from all of the conditional (posterior) densities associated with the joint posterior density of interest. As will be discussed in Section 4.4, as long as certain conditions are satisfied by both the joint posterior and the induced conditionals, these drawings represent a realization of a Markov chain with equilibrium distribution equal to the joint posterior. The drawings pertaining to any particular parameter also represent a simulated sample from the marginal posterior density of that parameter.

Demonstrating the procedure for the case of our specific parameter groupings: β, ϕ, $\mathbf{b}_{22}$ and Σ, the steps of the algorithm are as follows:

Step 1 Specify initial values for ϕ, $\mathbf{b}_{22}$ and Σ, $\phi^{(0)}$, $\mathbf{b}_{22}^{(0)}$ and $\Sigma^{(0)}$.

Step 2 Cycle iteratively through the four conditional densities, drawing respectively:

1. $\beta^{(i)}$ from $p_1(\beta^{(i)}|\phi^{(i-1)}, \mathbf{b}_{22}^{(i-1)}, \Sigma^{(i-1)}, \mathbf{y}, \mathbf{x})$,
2. $\phi^{(i)}$ from $p_2(\phi^{(i)}|\beta^{(i)}, \mathbf{b}_{22}^{(i-1)}, \Sigma^{(i-1)}, \mathbf{y}, \mathbf{x})$,
3. $\mathbf{b}_{22}^{(i)}$ from $p_3(\mathbf{b}_{22}^{(i)}|\beta^{(i)}, \phi^{(i)}, \Sigma^{(i-1)}, \mathbf{y}, \mathbf{x})$ and
4. $\Sigma^{(i)}$ from $p_4(\Sigma^{(i)}|\beta^{(i)}, \phi^{(i)}, \mathbf{b}_{22}^{(i)}, \mathbf{y}, \mathbf{x})$ until $i = M$.

Given the satisfaction of the required convergence conditions, the realized values, viewed as random variables, converge in distribution to the relevant marginal and joint distributions as $M \to \infty$. Alternatively, with M being large enough for convergence to have occurred, the continued application of the algorithm for a further N iterations produces both a sample of N $(\beta, \phi, \mathbf{b}_{22}, \Sigma)$ values from the joint posterior density and a sample of N values of any individual parameter (parameter set) from its marginal (joint) density.

Obviously, in order for the Gibbs Sampler to be operational, one needs to be able to sample from the conditional densities. The presence of both the conditional Jeffreys' prior for β given ϕ and $\mathbf{b}_{22}$ and the marginal Jeffreys' prior for ϕ render the conditional densities of the parameter vectors ϕ and $\mathbf{b}_{22}$ non-standard in form, such that direct simulation from them is not possible. Several different options are available in such a circumstance, all of which are variants on the idea of drawing from the unattainable distributions indirectly, via another distribution.. We choose to use the so-called Metropolis algorithm.

The reasons for the choice of the Metropolis algorithm are two-fold. First, being another Markov chain algorithm, if embedded appropriately within an outer Gibbs algorithm, the theory of Markov chains can be applied in a straight forward manner to the hybrid chain so produced to prove convergence. Second, we show that the nature of the identification problem is such that there is an obvious choice for the so-called candidate density to be used in the algorithm, one which, moreover, remains appropriate no matter what the dimension of ϕ and $\mathbf{b}_{22}$.

4.2 The Metropolis Algorithm

The Metropolis algorithm was originally proposed by Metropolis et al. (1953). Since then, a variety of different versions of the algorithm have been proposed and used within different statistical contexts. (See, for example, Hammersley and Handscomb (1964, Section 9.3) and Hastings (1970)). The Hastings version of the algorithm represents a sufficiently general one to nest certain other special cases, and is therefore a useful version to define. We shall demonstrate it with reference to generation from the conditional density for ϕ ($p_2(.)$ above). The description is directly applicable to generation from the conditional density of $\mathbf{b}_{22}$ also.

The basic thrust of the Metropolis algorithm is to simulate a value of ϕ indirectly, via a so-called candidate density q, with the latter having the properties of being both a good match for p_2 and easy to simulate from. Having a non-standard form, the integrating constant of p_2 is obviously unknown. Fortunately, the Metropolis algorithm uses p_2 in its unnormalized form only. Any reference to p_2 in this section will therefore be modulo integrating constant.

The steps of the Metropolis procedure, as inserted into the Gibbs algorithm at iteration i and as allowed to run itself for k iterations, are as follows:

Step 1 Given values $\beta^{(i)}$, $\mathbf{b}_{22}^{(i)}$ and $\Sigma^{(i)}$ as produced at the ith iteration of the Gibbs algorithm, parameterize a candidate density q via these values.

Step 2 Draw a candidate value for $\phi^{(i)}$, ϕ^*, from $q(\cdot|\beta^{(i)}, \mathbf{b}_{22}^{(i)}, \Sigma^{(i)})$

Step 3 Calculate the probability:

$$\begin{aligned}\alpha(\phi^{(i-1)}, \phi^*) &= \min\{(p_2(\phi^*)/q(\phi^*))/(p_2(\phi^{(i-1)})/q(\phi^{(i-1)})), 1\} \\ &\qquad if \quad p_2(\phi^{(i-1)})q(\phi^*) > 0; \\ &= 1 \\ &\qquad if \quad p_2(\phi^{(i-1)})q(\phi^*) = 0,\end{aligned}$$

where $\phi^{(i-1)}$, being the value for ϕ as produced in iteration $i-1$ of the Gibbs algorithm, represents the starting value for the Metropolis subchain in iteration i.

Step 4 With probability α, take ϕ^* as the first value for $\phi^{(i)}$ in the sub-chain of length k, say $\phi^{(i)}(1)$, otherwise take $\phi^{(i)}(1) = \phi^{(i-1)}$.

Step 5 Generate values for $\phi^{(i)}(j)$, $j = 2, 3, \ldots, k$, by cycling repeatedly through **Steps 2 to 4**, with the relevant comparison value ($\phi^{(i-1)}$ in the above description) being $\phi^{(i)}(j-1)$ at iteration j.

Step 6 Take $\phi^{(i)}(k)$ as the value for $\phi^{(i)}$.

On the assumption that both q and p_2 satisfy certain regularity conditions, continuation of this process produces a Markov chain with transition probabilities related to the above probability of acceptance and with equilibrium distribution equivalent to p_2. Subsequent to convergence to this distribution, any realization of the chain can be viewed as an observation from p_2.

Strictly speaking, the above steps would appear to produce what Tierney (1991) terms an independence chain, since the parameterization of q is not updated within the sub-chain to cater for the changing values in that chain. As noted in the Appendix, the use of such a chain enables stronger conclusions to be drawn with reference to convergence.

4.3 The Gibbs/Metropolis Strategy as Applied to the Cointegration Model

We shall now outline the precise form of the conditional densities which arise in the case of our particular model. Utilizing the notation of Section 3., and remembering that $\mathbf{x}^*$ and $\mathbf{x}^{**}$ are functions of the parameter vectors ϕ and $\mathbf{b}_{22}$ respectively, the joint posterior is proportional to:

$$\begin{aligned} &|\Sigma|^{-(n+3)/2} . \exp\{(-1/2) tr \Sigma^{-1} S\}. \\ &\{(\mathbf{x}^{*\prime}\mathbf{x}^{*}).(\mathbf{x}^{**\prime}\mathbf{x}^{**}) - (\mathbf{x}^{*\prime}\mathbf{x}^{**})^2\}^{1/2}.\{E \sum_t \left|\mathbf{u}^{r}_{1t-1}\mathbf{u}^{r\prime}_{1t-1}\right|\}^{1/2}, \end{aligned} \tag{15}$$

In applying the formal convergence criteria from Markov chain theory, it shall be convenient for us to eliminate from the support the sub-space on which the joint posterior density is rendered equal to zero by the conditional Jeffreys' prior. As such, from this point on we shall define the almost everywhere (with respect to Lebesgue measure) equivalent joint posterior as (15) defined on the support $D^* = D \cap \{(\beta, \phi, \mathbf{b}_{22}, \Sigma); \mathbf{x}^* \neq \mathbf{x}^{**}\}$, where $D = \mathbb{R}^1 \times \mathbb{R}^{p_{11}} \times \mathbb{R}^{p_{22}} \times \mathbb{S}^{pds}$.

4.3.1 $\beta|\phi, \mathbf{b}_{22}, \Sigma, \mathbf{y}, \mathbf{x}$

A little algebraic manipulation leads to a conditional density for β which is Normal in form, with mean, $\bar{\beta} = B_1/B_2$ and variance, $var(\beta) = \sigma_{11.2}B_2^{-1}$, where $B_1 = \sum_t(x_t^*[y_t^* - (\sigma_{12}/\sigma_{22})x_t^{**}])$, $B_2 = \sum_t(x_t^*)^2$ and:

$$\begin{aligned} y_t^* &= y_t - \phi_1 y_{t-1} - \phi_2(y_{t-1} - y_{t-2}) - \cdots - \phi_{p_{11}}(y_{t-p_{11}+1} - y_{t-p_{11}}) \\ &= y_t - \phi' \mathbf{y}_{t-1}^r \end{aligned}$$

From the form of the conditional density of β, we note that when conditioning on Σ is maintained, the identification problem which eventuates after integration with respect to Σ, does not present itself. We also note that the prior density, not being a function of β, does not impinge on the conditional density of β.

4.3.2 $\phi|\beta, \mathbf{b}_{22}, \Sigma, \mathbf{y}, \mathbf{x}$

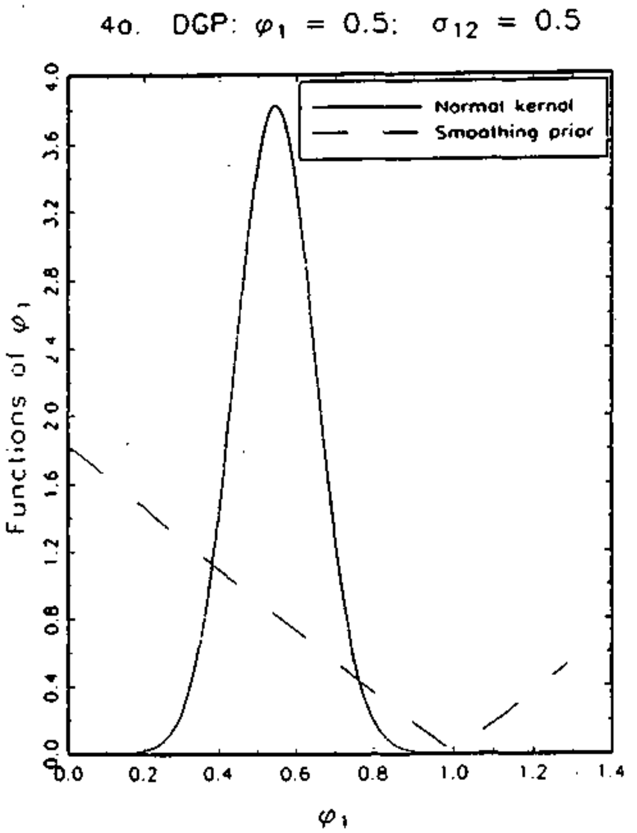
Without the presence of the prior density, similar manipulations to those used in the derivation of the density for β would lead to a multivariate Normal density for ϕ with mean, $\bar{\phi} = (U_1^{r\prime}U_1^r)^{-1}(U_1^{r\prime}[\mathbf{u}_1 - (\sigma_{11}/\sigma_{22})\mathbf{x}^{**}])$ and variance, $var(\phi) = \sigma_{11.2}(U_1^{r\prime}U_1^r)^{-1}$, where $U_1^r = (\mathbf{u}_{10}^r, \mathbf{u}_{11}^r, \ldots, \mathbf{u}_{1n-1}^r)'$.

With the inclusion of the prior density, the conditional density for ϕ is proportional to the product of the kernel of this Normal density and the full prior, viewed as a function of ϕ:

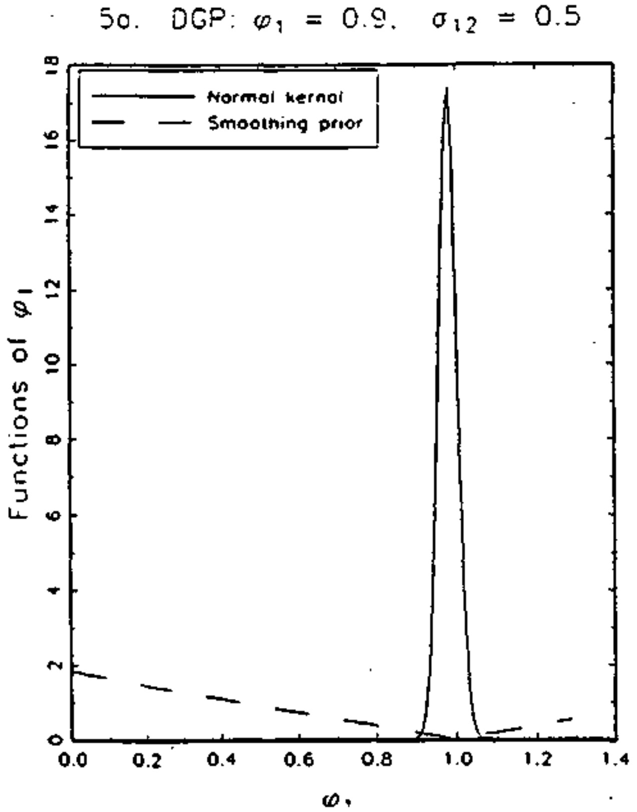
$$\{\sum_t(x_t - \phi'\mathbf{x}_{t-1}^r)^2 . \sum_t(x_t^{**})^2 - [\sum_t(x_t - \phi'\mathbf{x}_{t-1}^r)(x_t^{**})]^2\}^{1/2} . \{E\sum_t\left|\mathbf{u}_{1t-1}^r \mathbf{u}_{1t-1}^{r\prime}\right|\}^{1/2} \tag{16}$$

It is obvious that this product produces a conditional density which is non-standard. Let us assume for the moment that the final line in (16) is not a function of ϕ, so that the only complicating factor is the conditional smoothing prior. Due to the way in which ϕ enters this factor, conditioning on the remaining parameters does help to simplify the expression of the prior, viewed as a function of ϕ. The ordinate of the conditional density for is the product of the multivariate Normal ordinate and the ordinate of a function which possesses the multi-dimensional analogue of a v-shape. The latter function attains a minimum value which is arbitrarily close to zero at the set of values for the ϕ_i (arbitrarily close to the set) at which the identification problem arises.

Figures 4. and 5. Illustration of the impact of the smoothing prior at the conditional level


4a. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$

4b. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$


5a. DGP: $\varphi_1 = 0.9$. $\sigma_{12} = 0.5$

5b. DGP: $\varphi_1 = 0.9$; $\sigma_{12} = 0.5$

In the case of being one-dimensional (being denoted by ϕ_1), the situation can be represented graphically as in Figures 4 and 5. In Figures 4a and 5a we present, for dgp's with a true value for of 0.5 and 0.9 respectively, the Normal kernel and the smoothing function, both separately numerically normalized. In Figures 4b and 5b, the normalized product, or the conditional density of ϕ_1, is then presented along with the Normal density alone. The parameters on which the densities depend are assigned values via application of 20 iterations of the hybrid MCMC chain. Given that a single set of conditioning parameters are being used, the diagrams represent only one Normal kernel and conditional density from the whole set of such functions.

It is obvious in Figure 4, that as a result of the true dgp causing the Normal kernel to be centred well below one, the v-shaped smoothing function has minimal impact. As such, the overall conditional density for ϕ_1 is well matched by the Normal density. The Normal density also serves as a good match for the conditional density in Figure 5. We do notice, however, a small hump after 1 in the conditional density, obviously a consequence of the slightly more significant impact of the smoothing factor on the Normal kernel, when the latter has more mass in the region close to 1.

The point of this demonstration is twofold. First, it indicates the way in which the conditional Jeffreys' prior has only a relatively minor effect at the conditional level. This compares with its impact at the marginal level, where it serves to render a density which is very irregular, well-behaved. Second, it highlights the fact, related to the first, that the irregularity which obtains in the marginal density of ϕ_1 and for which the Jeffreys' prior factor has been introduced, is not present at the level of the conditional density. As such, that part of the conditional density which would be present if no adjustment were made for the identification problem, namely the Normal kernel, represents a well-behaved approximation to the overall conditional density which incorporates the smoothing factor.

Assuming that these same points are relevant when ϕ is of more than one dimension, the (normalized) Normal kernel, with mean and variance as given above would therefore appear to represent an obvious choice for candidate density q in a Metropolis algorithm as applied to the conditional density of ϕ.

When allowance is made for a non-constant function of ϕ in the last line of (16), the above reasoning would appear to continue to apply. A small amount of experimentation with different data sets indicates that the impact of the marginal Jeffreys' prior for ϕ_1 is minimal at the conditional level, at least when the Normal kernel does not have a significant amount of mass around 1. This is despite the fact that at the marginal level, it is sometimes the case that the marginal Jeffreys' prior on ϕ_1 induces bimodality in the posterior density.[7] We shall consequently continue to use the Normal kernel as candidate density, but with note being taken of the convergence behaviour of the algorithm, as compared with that of the algorithm incorporating a flat marginal prior on ϕ_1.

4.3.3 $\mathbf{b}_{22}|\beta, \phi, \Sigma$

The conditional density of the remaining set of autocorrelation parameters can be derived in similar fashion to that of ϕ. Define $\mathbf{u}_{2t-1} = (u_{2t-1}, u_{2t-2}, \ldots, u_{2t-p_{22}})'$, $U_2 = (\mathbf{u}_{20}, \mathbf{u}_{21}, \ldots, \mathbf{u}_{2n-1})'$, $\mathbf{u}_2 = (u_{21}, u_{22}, \ldots, u_{2n})'$ and $\sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$. The conditional density of $\mathbf{b}_{22}$ is proportional to:

$$\exp\{-1/(2\sigma_{22.1})(\mathbf{b}_{22} - \bar{\mathbf{b}}_{22})'(U_2'U_2)(\mathbf{b}_{22} - \bar{\mathbf{b}}_{22})\}.$$

$$\{\sum_t (x_t^*)^2 . \sum_t (x_t - (1, \mathbf{b}_{22}')\mathbf{x}_{t-1}^r)^2 - [\sum_t (x_t^*)(x_t - (1, \mathbf{b}_{22}')\mathbf{x}_{t-1}^r)]^2\}^{1/2}, \quad (17)$$

where the first part of the density is a Normal kernel, with mean vector and covariance matrix given respectively by $\bar{\mathbf{b}}_{22} = (U_2'U_2)^{-1}(U_2'[\mathbf{u}_2 - (\sigma_{12}/\sigma_{11})\mathbf{u}_1^*])$ and $var(\mathbf{b}_{22}) = \sigma_{22.1}(U_2'U_2)^{-1}$, with $\mathbf{u}_1^*$ being the vector of observations on $u_{1t}^* = u_{1t} - \phi'\mathbf{u}_{1t-1}^r$.

Since, the elements of $\mathbf{b}_{22}$ enter the smoothing factor of the prior density in exactly the same manner as do the elements of ϕ, the same comments which lead there to the choice of the Normal kernel as candidate density in the requisite application of a Metropolis algorithm, apply here.

4.3.4 $\Sigma|\beta, \phi, \mathbf{b}_{22}$

The conditional density for Σ can be read directly from the joint density as Inverted Wishart with n degrees of freedom and parameter matrix S^{-1}.

Generation of values from the Normal densities involves a straight forward use of the GAUSS command RNDN. Generation of values from the Inverted

---

[7] See Phillips (1991a) for a demonstration of a similar finding in the AR(1)/unit root model case.

Wishart density is achieved by taking the inverse of matrix values generated from the associated (non-inverted) Wishart density. The latter simulation is performed by generating (via RNDN) n two-dimensional normal deviates $\mathbf{z}_i = (z_{i1}, z_{i2})'$, with mean zero and variance covariance matrix S^{-1}. The random matrix $\Sigma^{-1} = \sum_i \mathbf{z}_i\mathbf{z}_i'$ then represents a realization of a Wishart variable with degrees of freedom n and covariance matrix S^{-1}. The inverse of this realization is the required realization of Σ.

Once the simulated values have been produced via the hybrid algorithm, estimates of the marginal densities of interest need to be produced for the purpose of inference. Via a Rao-Blackwell type argument, (see Gelfand and Smith (1990) for discussion on this point), it can be shown that the most accurate estimate of either marginal posterior of interest, $p(\beta|\mathbf{y},\mathbf{x})$ or $p(\phi_1|\mathbf{y},\mathbf{x})$, is a finite mixture density estimate. Demonstrated for the β marginal, this estimate is given by:

$$p(\hat{\beta|}\mathbf{y}) = (1/N)\sum_{i=1}^{N} p_1(\beta|\phi^{(i)}, \mathbf{b}_{22}^{(i)}, \Sigma^{(i)}, \mathbf{y}, \mathbf{x}), \tag{18}$$

where N is the number of simulated sets of parameter values. (18) is, of course, simply the sample estimate of the expectation implicit in the relationship between a conditional and a marginal density.

Since the conditional density for β is Normal in form, the component densities to be used in the mixture density estimate are known in their entirety; i.e. including their integrating constant.[8] In the case of ϕ_1 however, the relevant conditional density, which is simply the full conditional density for the vector ϕ but with only ϕ_1 viewed as the argument of interest, has a non-standard form. As a consequence, one-dimensional numerical integration needs to be performed on each of the N components in the mixture density estimate. This requirement obviously has implications for the speed with which results can be produced. However, the impact does not appear to be burdensome.

4.4 Convergence of the Markov Chain

There are two points to consider in relation to the question of convergence of the hybrid Markov chain to the joint posterior distribution. First, the

---

[8] If β had been a multidimensional vector, its conditional density would have had the multivariate Normal form. However, if only one particular element of β were of interest, the marginal density estimate would simply be based on the one-dimensional conditional density as read directly from the multivariate conditional.

structure of the Markov chain must be such that this joint posterior represents the so-called stationary or *invariant* distribution of the chain. Second, the structure of the chain must also be such that convergence towards the invariant distribution does indeed occur; i.e. that the chain is *ergodic*. If a straight Gibbs sampler were being used, then we would just need to consider these two points as they pertain to it alone. However, with a Metropolis algorithm embedded within the Gibbs Sampler at two points, we need to also ensure convergence to the relevant conditional distributions, so that the overall algorithm converges to the joint distribution.

4.4.1 Convergence of the Metropolis sub-chains

In the Metropolis sub-chains, we require convergence to the conditional distributions of and ϕ and $\mathbf{b}_{22}$ respectively. If it can be shown that the associated conditional densities are well-behaved bounded functions, then the distributions are both non-zero, finite probability measures over the space $(D^*, \mathbb{D}^*)$, where $\mathbb{D}^*$ is the σ-field countably generated from D^*. As per the discussion in the Appendix, the conditional distributions are, in that event, feasible invariant probability distributions for the Metropolis chains. In both applications of the Metropolis algorithm, the conditional density is the product of a Normal density kernel and the prior function. For $\mathbf{b}_{22}$, the latter comprises the $C_2^{1/2}$ function viewed as a function of $\mathbf{b}_{22}$, whilst for ϕ, the prior comprises the product of $C_2^{1/2}(\phi)$ and the marginal Jeffreys' prior on ϕ. As long as the tails of the Normal kernel dominate the prior function in each case, as the argument of the density moves towards the outer regions of D^*, then the density will be bounded. This issue needs to be examined further theoretically. However, numerical consideration of the situation, at least for the case of the reduced model with ϕ one-dimensional and $\mathbf{b}_{22} = 0$, suggests that such a density is bounded.

With reference to the results in the Appendix, we can say that the Metropolis sub-chains are *uniformly ergodic* for the relevant conditional distributions, denoted by $p(.)$, if the ratio $p(.)/q(.)$ in each case is both bounded and bounded away from zero. In each case, the density $p(.)$ is positive everywhere on D^* as a result of the definition of D^* and, as argued above, apparently bounded. Being a Normal density function, $q(.)$ is obviously both bounded and never equal to zero. As such, the ratio of $p(.)/q(.)$ can be assumed to satisfy the sufficient condition for uniform ergodicity.[9]

4.4.2 Convergence of the Gibbs chain

Assuming that at both points at which it is used, the Metropolis algorithm is run until convergence to the relevant conditional distribution is attained, the hybrid algorithm can be viewed as producing successive iterations from all of the full conditional distributions induced by the joint posterior. As indicated by the results in the Appendix, whether these iterations then represent successive values of a Markov chain with the joint posterior as equilibrium distribution, depends on the features of both the joint distribution and the induced conditional densities, where the latter comprise the transition kernel of the Gibbs algorithm.

In this case, the first step is to ascertain whether or not the joint posterior distribution defines a non-zero, finite measure on D^*. Once again, the latter requirement is determined by the nature of the interaction between the prior and the remaining part of the posterior function. We assume that this interaction is such that the total mass of the joint measure on D^* is finite and the joint posterior a valid invariant probability distribution as a consequence.

Sufficient conditions for convergence to the joint posterior are that the latter is lower semicontinuous at zero, that the integral of the joint density with respect to each individual parameter set is locally bounded and that D^* is connected. All of these conditions are satisfied in our particular situation. As such, we can conclude that the outer Gibbs chain is *(simply) ergodic.*

5 Numerical Application of the Method

We shall now apply the proposed method to particular sets of numerical data. The data is all artificially simulated, with a view to demonstrating both the forms of the relevant densities given different underlying generating processes, and the nature of the convergence of the MCMC method in different environments. In order to get some feel for the repeated sampling performance of inferences based on the marginal densities, in comparison

---

[9] We note that since D^* has the subspace on which $p(.) = 0$ omitted, we shall essentially restrict the domain of $q(.)$ correspondingly. This amounts to simply ignoring any values drawn from $q(.)$ which fall into this subspace, and drawing another set for entry into the algorithm.

with relevant Classical alternatives, we present the results of a small Monte Carlo experiment.

The fundamentals of our method are quite adequately illustrated within the context of the very simple framework based on $b_{11}(L) = 1 - \phi_1 L$ and $b_{22}(L) = 1$. An added advantage of using such a simple specification is that the marginal densities of ϕ_1 and β can be easily produced via low-dimensional numerical integration, for the purpose of illustrating the accuracy of the MCMC method in reproducing the "exact" densities.

The model used to generate the data is thus:

$$\begin{aligned} y_t &= \beta x_t + u_{1t}, \\ x_t &= x_{t-1} + u_{2t}, \\ u_{1t} &= \phi_1 u_{1t-1} + e_{1t} \qquad \text{and} \\ u_{2t} &= e_{2t}, \end{aligned} \tag{19}$$

with $(e_{1t}, e_{2t})'$ generated as multivariate Normal with zero mean and variance covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \qquad , \sigma_{12} = \sigma_{21}.$$

In Sections 5.1 and 5.2, we illustrate the impact on the marginal densities of ϕ_1 and β respectively, and the MCMC estimation thereof, of variation in ϕ_1 in the true dgp. The values of β, σ_{12}, σ_{11} and σ_{22} in the generating process remain fixed at 3, 0.5, 1 and 1 respectively. As a consequence of the latter two values, σ_{12} can be interpreted as the correlation between e_{1t} and e_{2t}. The MCMC density estimates are produced using the following iteration strategy. After a "burn-out" period of M, we take into the sample the output of every rth iteration, the intermittent sampling of the chain aimed at speeding up convergence via a breaking of the Markovian dependence. With N denoting the final number of sample values from which density estimates are constructed, the MCMC densities are produced from a total of $(r + [N + (M/r)]) - M$ iterations for the outer Gibbs chain. For all examples, we use $M = 400$ and $r = 10$. For a reasonable degree of convergence to be attained, we require different values of N. Typically, values ranging from 200 to 500 for N are sufficient. In all instances, we perform 20 iterations of the Metropolis sub-chain before taking a value as a realization from the relevant conditional density. Experimentation with the number of Metropolis iterations here indicated that little was gained, in terms of the accuracy of the final density estimates, by increasing the number beyond 20.

Where it is appropriate, we provide with each graph, the appropriate summary of the relevant density. In the case of the marginal ϕ_1 densities for example, we provide the mode of the density, plus the probability of either cointegration or non-cointegration, depending on the nature of the underlying dgp. In the case of the marginal β densities, we report the mode of each density. The reason why we concentrate upon modal point estimates is that we have yet to formally establish the existence of moments for the marginal densities.

5.1 The impact of true ϕ_1 on the marginal density of ϕ_1.

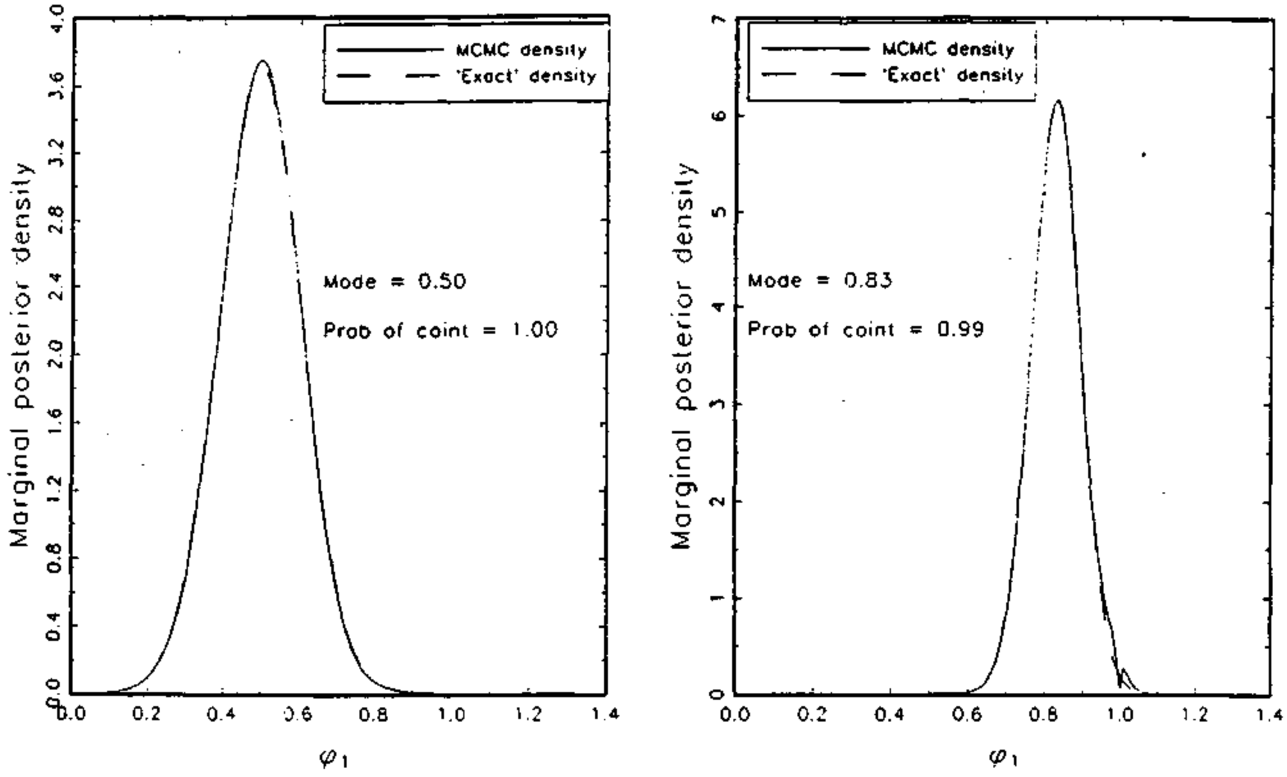
Figures 6, 7, and 8 demonstrate the types of marginal densities available for the computation of the relative probabilities of stationarity and non-stationarity. Figures 6 and 8a are based on a flat marginal prior for ϕ_1 whilst Figures 7 and 8b use a Jeffreys' marginal prior. The inferences yielded by of all densities in Figures 6 and 7 are very accurate. When the true ϕ_1 is less than 1 and cointegration is present as a consequence, the densities assign probabilities to that hypothesis which exceed 0.97. With the sample produced from an underlying dgp with $\phi_1 = 1$, as in Figure 8, we see that, as anticipated, the marginal Jeffreys' prior produces a larger probability content in the non-stationary region than does the flat marginal prior. In the Monte Carlo results presented below, we see that this probability is close to 50% on average in repeated samples. We note that the modes of the marginal densities provide very accurate point estimates of the true ϕ_1 in all of the diagrams.

The MCMC densities reproduce the exact densities with great precision in Figures 6 and 7. In Figure 8a, convergence has virtually occurred, but with a large number of iterations having been required ($N = 1000$). In Figure 8b, convergence has not quite yet occurred, even with the larger number of iterations. As previously discussed, when the Normal candidate densities used in the Metropolis sub-chains have a lot of mass around 1, the impact of the prior function, particularly the marginal Jeffreys' prior, may be quite significant. Such a situation is obviously likely when the true $\phi_1 = 1$. In such a case, a very large number of iterations may be required for convergence.

Figures 6. and 7. Smoothed marginal posteriors for ϕ_1 based on a marginal and Jeffreys' prior for ϕ_1 respectively.

6a. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$

6b. DGP: $\varphi_1 = 0.9$; $\sigma_{12} = 0.5$



7a. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$

7b. DGP: $\varphi_1 = 0.9$; $\sigma_{12} = 0.5$

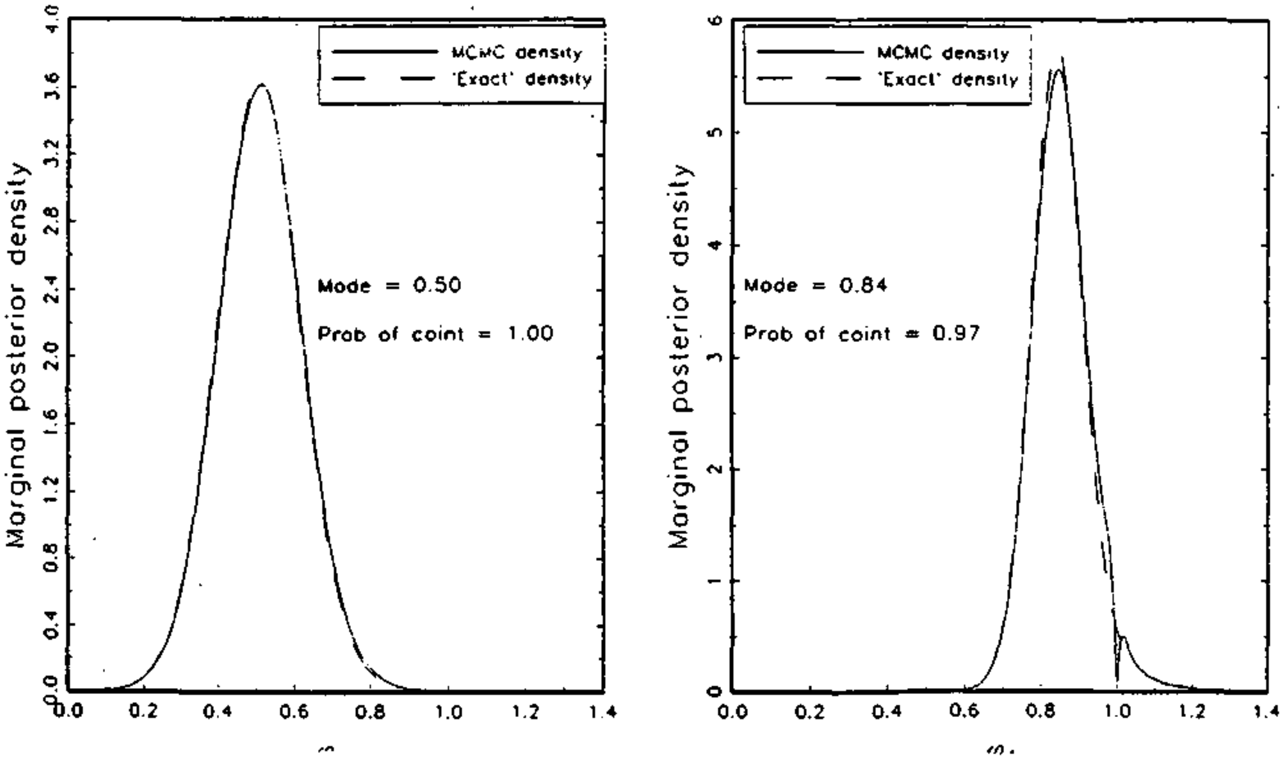
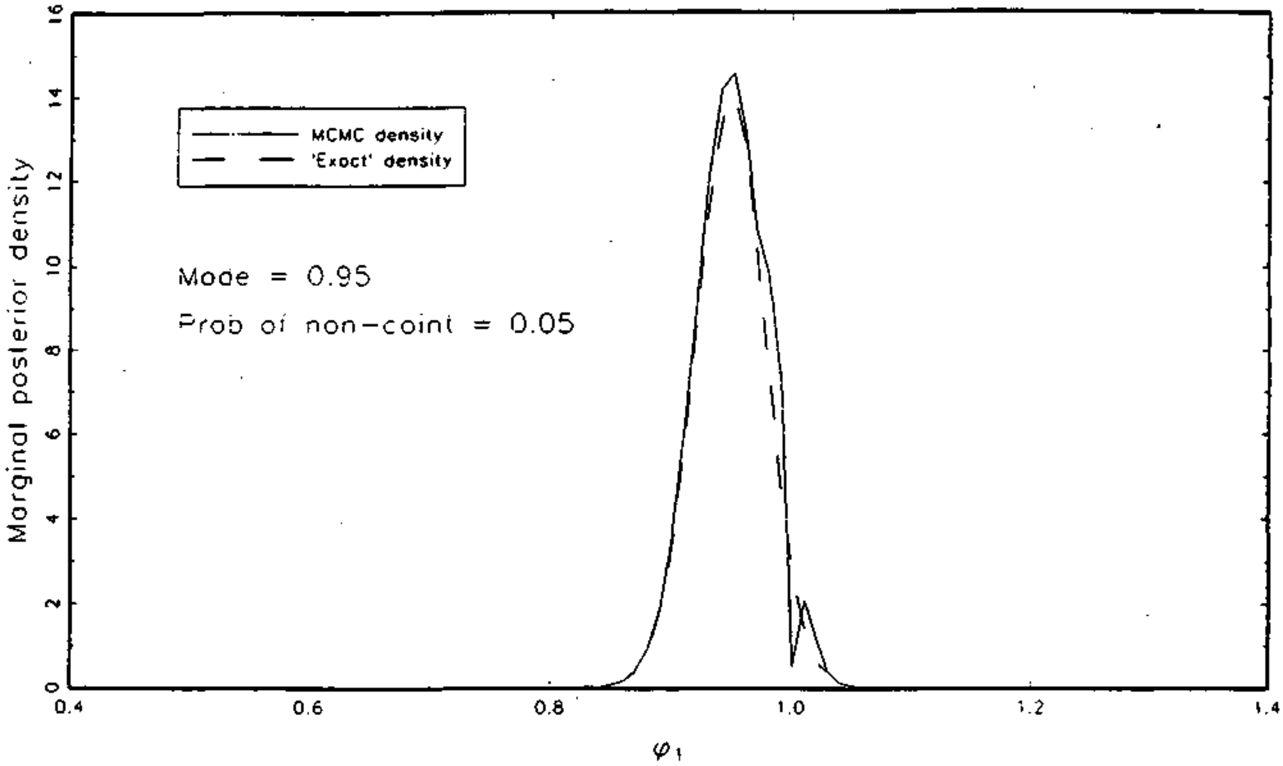


Figure 8. Smoothed marginal posteriors for ϕ_1 when the true $\phi_1 = 1$
(flat and Jeffreys' marginal prior respectively)

8a. DGP: $\varphi_1 = 1.0$; $\sigma_{12} = 0.5$



8b. DGP: $\varphi_1 = 1.0$; $\sigma_{12} = 0.5$

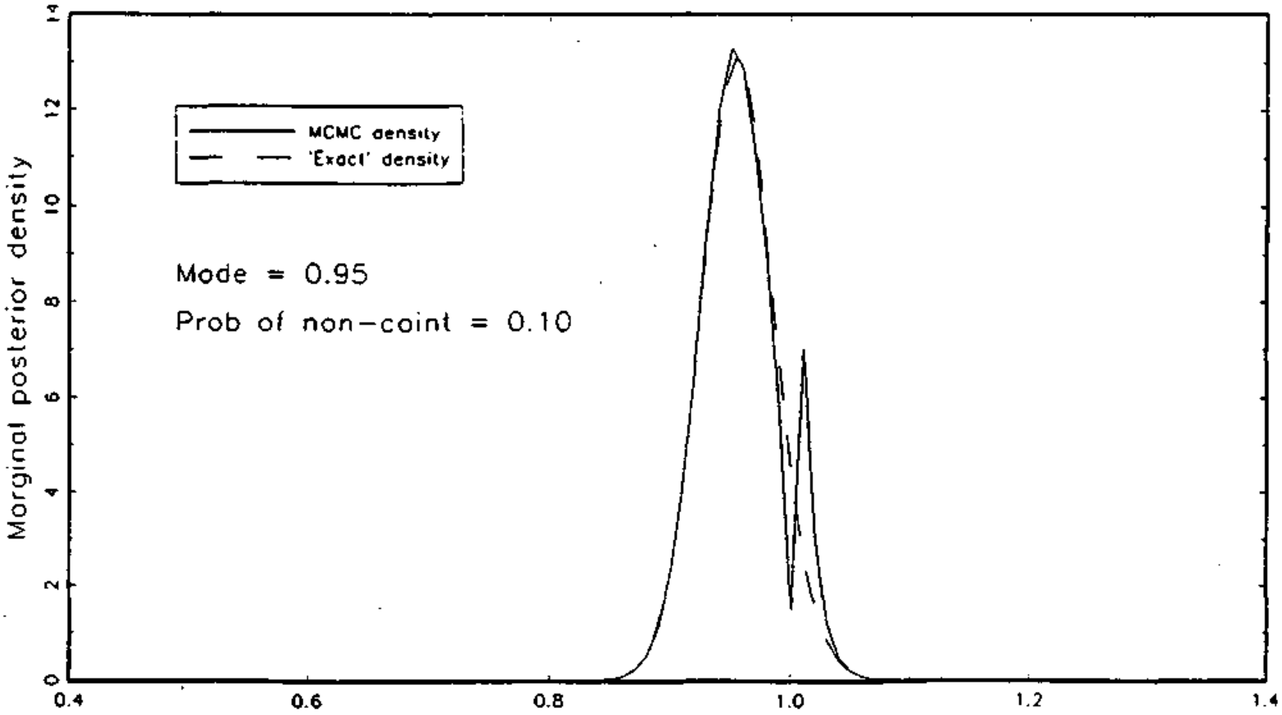
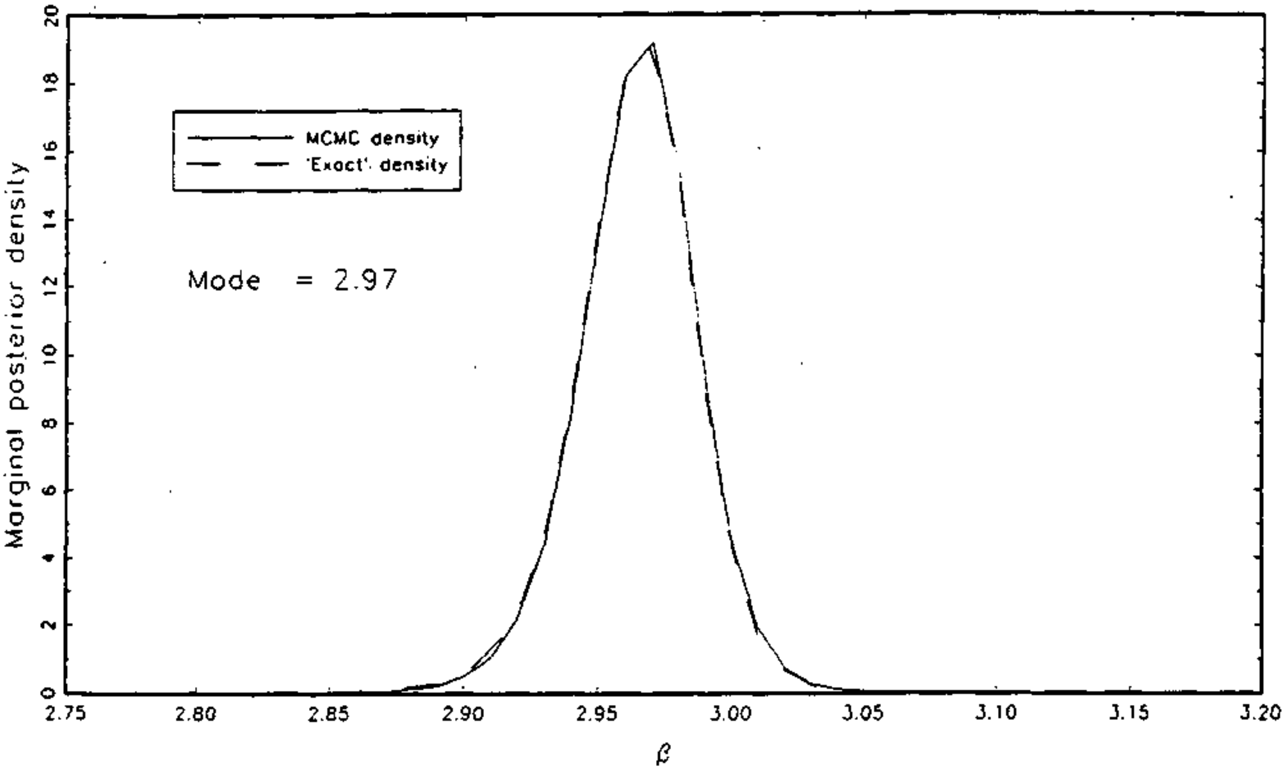
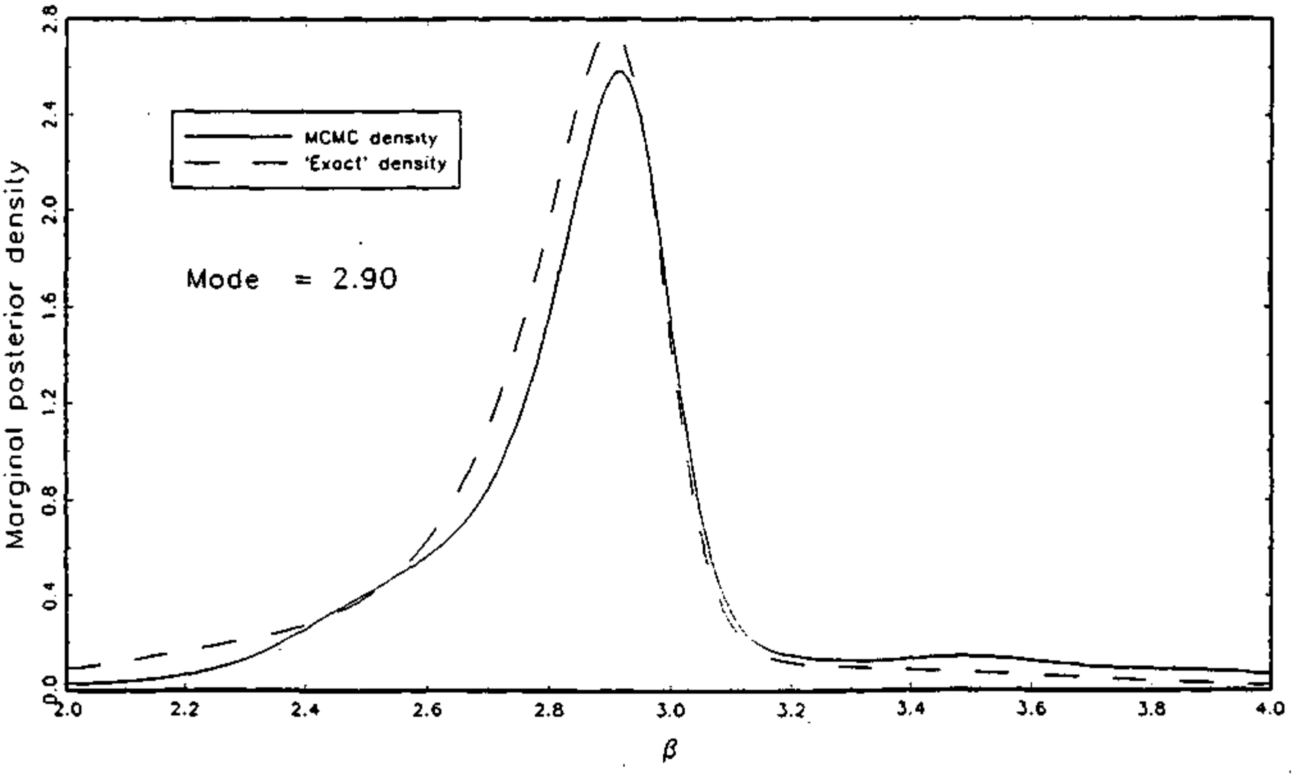


Figure 9. Marginal posteriors for β based on the conditional Jeffreys' prior

9a. DGP: $\varphi_1 = 0.5$; $\sigma_{12} = 0.5$; $\beta = 3$;



9b. DGP: $\varphi_1 = 1.0$; $\sigma_{12} = 0.5$; $\beta = 3$;



5.2 The impact of true ϕ_1 on the marginal density of β

We provide examples in Figures 9a and 9b of the types of densities which result when y_t and x_t are and are not cointegrated respectively.[10] A couple of points are worth noting. First, the MCMC density is able to accurately mimic the exact density in the cointegration case; less so in the non-cointegration case. Second, the density in the latter case provides the sort of information which one would wish for in this context: the density is very variable, making for a very wide interval estimate, with the tail behaviour indicating that moments may not exist. When x_t and y_t are not cointegrated, it is desirable that inference about β be imprecise in this way. When cointegration exists, on the other hand, inference is accurate both in terms of the location and the degree of dispersion of the marginal density. The Monte Carlo results provided in Section 5.3 demonstrate that this accuracy continues to obtain in a repeated sampling context, when the mode is used as a point estimator.

5.3 A Monte Carlo Experiment

The main aims of this very limited Monte Carlo study are twofold. First, we wish to assess the accuracy, in terms of bias and root mean square error (RMSE) in repeated samples, of our modal estimator of β compared with two of the main single equation Classical alternatives, namely the Fully Modified OLS procedure (FMOLS) of Phillips and Hansen (1990) and the method of Phillips and Loretan (1991) discussed earlier (PL). Second, we wish to compare the quality of our inferences regarding ϕ_1 and, hence, the presence of cointegration, with that of Classical inference based on the application of the Dickey Fuller test to the OLS residuals.

With reference to the second aim, we compute, in the case of a true stationary error, the average in repeated samples of the probability of cointegration as based on the two different possible marginal posteriors for ϕ_1; i.e. as based on a flat and Jeffreys' marginal prior respectively for ϕ_1. We then compare the nature of the information conveyed by this probability with the empirical power of the Dickey Fuller test. Note that we are not directly

---

[10] The marginal density of β is invariant to the marginal prior density used for ϕ_1.

comparing the two different probabilities, Bayesian and Classical, as they are not commensurate measurements. We are simply attempting to answer the question: given the true nature of the process underlying the data, which inferential method, whose properties are summarized by the two respective probabilities, is likely to provide the most accurate reflection of that process.

In the case of a true non-stationary error, we compute the average probability of non-cointegration, along with the empirical size of the Dickey Fuller test.[11]

The results for inference regarding β are given in Tables 1. The results for inference regarding ϕ_1 are given in Tables 2 and 3. The sample size used in all experiments was 50 and the number of replications 1000.

Table 1. Estimation of β

		$\phi_1 = 0.9$		$\phi_1 = 1.0$
		$\sigma_{12} = 0.5$	$\sigma_{12} = 0.9$	$\sigma_{12} = 0.5$
BIAS	*β MODE*	0.042	0.039	0.532
	FMOLS[12]	0.074	0.119	0.608
	PL	-2.080	-0.610	-5.516
RMSE	*β MODE*	0.244	0.150	1.065
	FMOLS	0.257	0.224	1.713
	PL	10.426	4.346	27.058

With reference to estimation of β when it is the parameter of a cointegrating relationship, the results indicate that the modal estimator has both smaller bias and smaller RMSE than the main Classical alternatives. Moreover, an increase in the correlation between the two underlying error processes, which is one aspect of the degree of endogeneity of x_t, does not appear to worsen the repeated sample behaviour of the mode, in contrast to the

---

[11] A 5% critical value of -2.9 for the Dickey Fuller test was computed numerically from 10,000 replications of the model under the null of $\phi_1 = 1$.

[12] The FMOLS procedure was implemented with the non-parametric estimate of the long-run variance based on a lag length appropriate for an AR(1) process with a parameter value as specified in the true dgp.

FMOLS procedure. When x_t and y_t are not cointegrated, on the other hand, the modal estimate, whilst still exhibiting an ability to estimate the value of more accurately than the FMOLS estimate, mimics in a broad sense the behaviour of the Classical estimator. That is, both are (relative to the cointegration case) imprecise estimators of β. Phillips (1986) has outlined the behaviour of various sample moments and certain statistics which are built from these moments, in such a context. For example, the OLS estimator converges to a random variable in the limit. The point is that the essence of these results is likely to be applicable to all the estimators considered here, in which case we would anticipate the highly variable results across samples which we have observed.

As regards the PL estimates, which have been computed using the Maximum Likelihood (ML) routine in GAUSS, we have found them to be extremely sensitive to the choice of starting values. In particular, any departure in the starting values from the true specification has been found to adversely affect the accuracy of the estimates. The tabulated results exemplify this situation. In contrast, our MCMC method, which also requires the specification of starting values, is consistent with the theoretical Markov chain property of being invariant to the particular choice made, as long as a sufficient number of iterations are performed.

The sensitivity of the ML estimates is presumably a reflection of the potential multicollinearity/identification problem associated with the PL model specification, which was alluded to earlier. Since we have adjusted for this problem, prior to applying the Markov chain procedure, we would anticipate the apparent insensitivity to starting values which the latter exhibits. We assume that these same comments would apply had NLS been used rather than ML estimation to produce the PL results.

As concerns inference about ϕ_1, the contrast between the Bayesian and Classical results is rather marked. The notorious lack of power of the residual-based tests is exemplified by the Dickey Fuller results. Even when the error term is clearly stationary, ϕ_1 equalling 0.9, the test has an empirical power of only 15% when $\sigma_{12} = 0.5$, less when σ_{12} is higher. The Bayesian method, on the other hand has an average probability of stationarity which is very high in all settings. Due to a tendency of the marginal ϕ_1 density, under both the flat and Jeffreys' marginal priors for ϕ_1, to become more concentrated about the true value of ϕ_1 as σ_{12} increases, the probability of stationarity tends towards a desirable figure of 100% as σ_{12} increases.

Table 2. Inference Regarding Cointegration

		$\phi_1 = 0.9$		$\phi_1 = 1.0$
		$\sigma_{12} = 0.5$	$\sigma_{12} = 0.9$	$\sigma_{12} = 0.5$
Average $\Pr(\phi_1 < 1)$	*Flat*	0.967	0.991	
	Jeffreys'	0.788	0.969	
Average $\Pr(\phi_1 \geq 1)$	*Flat*			0.229
	Jeffreys'			0.512
Emp. Power	*DF*	0.145	0.121	
Emp. Size	*DF*			0.058

When cointegration is not present, the Bayesian method, in particular as based on the marginal Jeffreys' prior, still gives a large probability, on average, to the correct hypothesis. Of course, since we are computing the probability of non-stationarity in the error (rather than the probability of a unit root per se), this probability would be larger, the further above 1 is the true value of ϕ_1, since the marginal densities would tend to be centred to the right of 1.

The impact on the modal estimates of ϕ_1 of the marginal Jeffreys' prior for ϕ_1 is interesting in the light of recent theoretical developments regarding the reduction in the bias of ML estimates obtained by "modifying" the likelihood function by a Jeffreys' prior.[13] As Table 5 indicates, there is a significant reduction in the small sample bias of the modal estimate of ϕ_1 when the marginal Jeffreys' prior on ϕ_1 is used. This is suggestive of the possibility that the theoretical results alluded to, which have been developed within the context of identically and independently distributed data, are also applicable

---

[13] See Firth (1993). A related strand of the literature, dating from the work of Welsh and Peers (1965) demonstrates the improvement in frequentist coverage of Bayesian interval estimates obtained using a Jeffreys' prior. See also the relevant discussion in Phillips (1991a and b). The way in which a Jeffreys' prior serves to produce inferences which tally, in some sense, with associated Classical inferences, can be viewed as another manifestation of its noninformativeness.

to dependent, possibly non-stationary data.

Table 3 Estimation of ϕ_1

		$\phi_1 = 0.9$		$\phi_1 = 1.0$
		$\sigma_{12} = 0.5$	$\sigma_{12} = 0.9$	$\sigma_{12} = 0.5$
BIAS	*Flat*	-0.061	-0.021	-0.055
	Jeffreys'	-0.023	-0.014	-0.018
RMSE	*Flat*	0.102	0.050	0.088
	Jeffreys'	0.134	0.051	0.099

6 Conclusions

The paper has presented a new way of approaching inference in a cointegration context. The inferences are based upon marginal posterior density functions, which are able to be accurately estimated by a combination of MCMC methods. The main contributions of the paper are twofold. First, it provides evidence that such posterior based inference may provide highly informative and accurate information about both the presence of cointegration and the nature of the cointegrating relationship. Second, it provides an easy to implement strategy for accurately estimating the relevant marginal densities in the typical case where the dimension of the model precludes numerical integration.

References

[1] Box, G.E.P. and G.C. Tiao, 1973, *Bayesian inference in statistical analysis*, (Addison-Wesley Publishing Co., Reading).

[2] Casella, G. and E.I. George, 1992, Explaining the Gibbs sampler, *The American Statistician* 46, 167-174.

[3] Chib, S., 1993, Bayes regression with autoregressive errors: A Gibbs sampling approach, *Journal of Econometrics* 58, 275-294.

[4] Chib, S. and E. Greenburg, 1993, Posterior Analysis of SUR models via Markov chain Monte Carlo, *Working Paper No. 174, Washington University in St. Louis.*

[5] Engle, R.F.,D.F. Hendry and J-F. Richard, 1983, Exogeneity, *Econometrica* 51, 277-304.

[6] Firth, D., 1993, Bias reduction of maximum estimates *Biometrika* 80, 27-38.

[7] Gelfand, A.E. and A.F.M. Smith, 1990, Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association* 85, 398-409.

[8] Hastings, W.K., 1970, Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97-109.

[9] Hammersley and Handscomb, 1964, *Monte Carlo methods*, (Methuen, London).

[10] Kleibergen, F., and H.K. van Dijk, 1994a, On the shape of the likelihood/posterior in cointegration models, *Econometric Theory* 10, 514-551.

[11] Kleibergen, F., and H.K. van Dijk, 1994b, Bayesian analysis of simultaneous equation models using noninformative priors, *Working paper, Econometric Institute and Tinbergen Institute, Rotterdam, The Netherlands.*

[12] Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, 1953, Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087-1092.

[13] Phillips, P.C.B., 1986, Understanding spurious regressions in Econometrics, *Journal of Econometrics* 33, 311-340.

[14] Phillips, P.C.B., 1991a, To criticize the critics: an objective Bayesian analysis of stochastic trends, *Journal of Applied Econometrics* 6, 333-364.

[15] Phillips, P.C.B., 1991b, Bayesian routes and unit roots: de rebus prioribus semper est disputandum, *Journal of Applied Econometrics* 6, 435-474.

[16] Phillips, P.C.B., 1991c, Optimal inference in cointegrated systems. *Econometrica* 59, 283-306.

[17] Phillips, P.C.B., 1993, The long-run Australian consumption function re-examined: an empirical exercise in Bayesian inference, *Cowles Foundation Paper No. 825.*

[18] Phillips, P.C.B., 1994, Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models, *Econometrica* 62, 73-93.

[19] Phillips, P.C.B., and B.E. Hansen, 1990, Statistical inference in instrumental variables regression with I(1) processes, *Review of Economic Studies* 57, 99-125.

[20] Phillips, P.C.B. and Loretan, M., 1991, Estimating long-run Economic equilibria, *Review of Economic Studies* 58, 407-436.

[21] Roberts, G.O. and A.F.M. Smith, 1994, Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms, *Stochastic Processes and their Applications* 49, 207-216.

[22] Schotman, P. and H.K. van Dijk, 1991, On Bayesian routes to unit roots, *Journal of Applied Econometrics* 6, 387-401.

[23] Smith, A.F.M. and G.O. Roberts, 1993, Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, B* 55, 3-23.

[24] Tanner, M.A., 1994, *Tools for statistical inference,* second edition (Springer-Verlag, New York, NY).

[25] Tierney, L., 1991, Markov chains for exploring posterior distributions, *Technical Report No. 560., University of Minnesota School of Statistics.*

[26] Welsh, B.L. and H.W. Peers, 1963, On formulae for confidence points based on integrals of weighted likelihoods, *Journal of the Royal Statistical Society, B* 25, 318-29.

A General Convergence Conditions for Markov Chains

We present the formal convergence theory as it relates both to general Markov chains and to the specific Gibbs and Metropolis algorithms. The papers to which we refer are Tierney (1991) and Roberts and Smith (1994). In addition, we cite the book by Nummelin (1984), which outlines the properties of Markov chains as defined on general measurable spaces.

A *time-homogenous (discrete time, continuous state space) Markov chain* on a measurable space $(E, \mathbb{E})$, where $\mathbb{E}$ is a σ-field countably generated from E, is a sequence of (possibly vector valued) random variables $\{X_n, n \geq 0\}$ on $(E, \mathbb{E})$ whose conditional, or transition, probabilities, satisfy the so-called Markov property and are independent of time. For such a chain, we can define the *transition kernal* on $(E, \mathbb{E})$ as $K : E \times E \to \mathbb{R}^+$, such that, with respect to a σ-finite measure ν on E:

$$K(X_0, A) = P\{X_{n+1} \in A | X_0, X_1, \ldots, X_n\} = P\{X_{n+1} \in A | X_n\} = P\{X_1 \in A | X_0\}$$

for all measurable A. Assuming that, for any $X_0 \in E$, K is absolutely continuous with respect to ν, $K(.,.)$ can be expressed as the integral of a *transition density kernal* k as follows:

$$K(X_0, A) = \int_A k(X_0, X_1) \partial\nu(X_1).$$

Recalling the use of the notation $p(.)$, $q(.)$ and $\alpha(.)$ for the target (invariant) density, candidate density and acceptance probability respectively for the Metropolis algorithm, the *Metropolis* density kernal (describing only accepted iterations) is defined as:

$$k_H(X_0, X_1) = q(X_0, X_1)\alpha(X_0, X_1).$$

The *Gibbs* transition density kernal is given by:

$$k_G(X_0, X_1) = p(X_1^{(1)}|X_2^{(0)}, X_3^{(0)}, \ldots, X_k^{(0)}).p(X_2^{(1)}|X_1^{(1)}, X_3^{(0)}, \ldots, X_k^{(0)}).\ldots$$
$$\ldots.p(X_k^{(1)}|X_1^{(1)}, X_2^{(1)}, \ldots. X_{k-1}^{(1)}),$$

where $X = (X_1, X_2, \ldots, X_k)$ defines the chosen blocking of an m-dimensional random variable X, with $1 < k \leq m$.

A Markov chain has an *invariant distribution*, π, if the transition kernal K satisfies:

$$\pi(A) = \int K(X_0, A)d\pi(X_0).$$

If $0 < \pi(E) < \infty$, then with the satisfaction of a convergence condition to be described below, π is the *unique* invariant probability distribution for the chain. (See Nummelin (1984, Chp. 5).

Having established that π is a positive, finite measure on the space, we then need to determine:

1. That the structure of $K(X_0, A)$ is such that π *is* the invariant distribution of the chain,; and

2. That π is the *equilibrium* distribution, in the sense that it is the unique invariant distribution *towards which convergence occurs* for π-almost all X_0 ($X_0 \in E$). This can be expressed as :

$$\lim_{n \to \infty} K^n(X_0, A) = \pi(A)$$

for π-almost X_0 and all measurable sets A. $K^n(.)$ denotes the nth iterate of the kernal K, representing the probability of entering set A after n steps of the chain from a starting value of X_0.

Assuming that **1.** is satisfied, **2.** requires that the Markov chain is both *irreducible* and *aperiodic*. A Markov chain with invariant distribution π is said to be *π-irreducible* if for each $A \in \mathbb{E}$ with $\pi(A) > 0$, $K^n(X_0, A) > 0$ for some $n \geq 1$. That is, the chain has a positive probability of entering a set A, to which the distribution π ascribes positive probability, in a finite number of steps from an initial value X_0. A chain is *periodic* if there exists an integer $d \geq 2$ and a sequence $\{E_0, E_1, \ldots, E_{d-1}\}$ of d non-empty disjoint sets in $\mathbb{E}$ such that for all $i = 0, 1, \ldots, d-1$, and all $X_0 \in E_i$, $P(X_0, E_j) = 1$ for $j = (i+1)(\bmod d)$. The chain is *aperiodic* if it is not periodic.

A π-irreducible Markov chain with unique invariant probability distribution π is *positive recurrent* in the sense that for each A with $\pi(A) > 0$:

$$\begin{aligned}&P\{X_n \in A \text{ } i.o.|X_0\} > 0 \text{ for all } X_0 \text{ and}\\&P\{X_n \in A \text{ } i.o.|X_0\} = 1 \text{ for } \pi\text{-almost all } X_0.\end{aligned}$$

The chain is *Harris recurrent* if $P\{X_n \in A \text{ } i.o.|X_0\} = 1$ for all X_0. The notation "*i.o.*" denotes infinitely often. (See Nummelin (1984, Chp.3)).

We can now bring together the above definitions and discussion into the following formal theorem, whose proof can be found in Nummelin (1984).

Theorem 1 *If a Markov chain $\{X_n, n \geq 0\}$ with kernal K has invariant (probability) distribution π and is π-irreducible, then π is the unique invariant distribution for $\{X_n\}$. If $\{X_n\}$ is also aperiodic then for π-almost X_0:*

$$\|K_n(X_0, .) - \pi\| \to 0$$

where $\|.\|$ denotes total variation distance. If $\{X_n\}$ is Harris recurrent, then the convergence occurs for all X_0. In the latter case, the chain is said to be ***simply ergodic.***

With reference to the *Metropolis* algorithm, we can state the following Lemmas, whose proofs are all to be found in Tierney (1991):

Lemma 1 *(a) For any q, a Metropolis chain is* ***reversible*** *in the sense that:*

$$p(X_0)q(X_0, X_1)\alpha(X_0, X_1) = p(X_1)q(X_1, X_0)\alpha(X_1, X_0).$$

This property ensures that the distribution P associated with the density p is the invariant *distribution of the chain.*

(b) A P-irreducible Metropolis chain is also Harris recurrent.

As a consequence of *Lemma 1.*, we can conclude that any P-irreducible, aperiodic Metropolis chain is *simply ergodic.*

Lemma 2 *An* independence *Metropolis chain with invariant distribution P is P-irreducible and aperiodic if and only if $q(.)$ is positive ν-almost everywhere on $\mathbb{E}$.*

For a independence chain, we can in fact go further, via the following Lemma:

Lemma 3 *An* ***independence*** *Metropolis chain is* ***uniformly ergodic*** *for P if $p(.)/q(.)$ is bounded and bounded away from zero, where an ergodic chain with kernal K and invariant distribution P is called uniformly ergodic if:*

$$\|K^n(X_0,.) - P(.)\| \leq Mr^n$$

for some $r < 1$ and constant M.

With reference to the *Gibbs* algorithm, conditional probability manipulations reveal that for well-defined conditional densities, $k_G(X_0, X_1)$ is associated with a kernal function which has invariant distribution equal to the distribution of X. The following Lemmas then provide conditions for the existence of a well-defined kernal density and for convergence to the invariant distribution (π). *Lemma 4.* requires the following definition:

Definition 1 *A function $h : \mathbb{R}^m \rightarrow \mathbb{R}^+$ is lower semicontinuous at 0 if, for all X with $h(X) > 0$, there exists an open neighbourhood $N_X \ni X$ and $\varepsilon > 0$ such that, for all $Y \in N_X$, $h(Y) \geq \varepsilon > 0$.*

Lemma 4 *For ν m-dimensional Lebesgue measure, if π is lower semicontinuous at zero, then $\int \pi(X)dX_j > 0$ for $j = 1, 2, \ldots, k$.*

Proof*: See Roberts and Smith (1994).*

If the condition in *Lemma 4* holds, then k_G is a well-defined kernal density and π an *invariant* distribution for the chain.

Lemma 5 *For ν m-dimensional Lebesgue measure, if π is lower semicontinuous at zero, E is connected and $\int \pi(X)dX_j$ is locally bounded, then the Gibbs chain is π-irreducible and aperiodic.*

Proof*: See Roberts and Smith (1994).*

Lemma 6 *Suppose a Markov chain with kernal K has invariant distribution π and is π-irreducible. If $K(X_{0,}.)$ is absolutely continuous with respect to π for all X_0, then $\{X_n\}$ is Harris recurrent.*

Proof: See Tierney (1991) and Nummelin (1984).

For standard problems, in which the multivariate space E can be decomposed as $\prod_i^k E_i$, it is straight forward to show that $\pi(A) = 0 \Rightarrow K(X_0, A) = 0$ for measurable sets A, when $K(.,.)$ is the Gibbs kernal. As such, Harris recurrence can be viewed as being satisfied by most Gibbs Samplers and the conditions in *Lemma 5.* viewed as sufficient for a Gibbs chain to be *simply ergodic* for π.