University of Leicester

DEPARTMENT OF ECONOMICS

BAYESIAN APPROACHES TO COINTEGRATION

Gary Koop, University of Leicester

Rodney Strachan, University of Leicester

Herman van Dijk, Erasmus University

Mattias Villani, Sveriges Riksbank and Stockholm University

Working Paper No. 04/27
September 2004

Bayesian Approaches to Cointegration

Gary Koop
Department of Economics
University of Leicester
Gary.Koop@leicester.ac.uk

Rodney Strachan
Department of Economics
University of Keele
r.w.strachan@econ.keele.ac.uk

Herman van Dijk
Econometric Institute
Erasmus University
hkvandijk@few.eur.nl

Mattias Villani
Research Department
Sveriges Riksbank
and
Department of Statistics
Stockholm University
mattias.villani@riksbank.se

1 Introduction

Ever since the seminal work of Nelson and Plosser (1982), common wisdom has held that many macroeconomic and financial time series contain stochastic trends (such as a random walk process) or unit roots. When the stochastic trends in various time series are independent of one another, then no stable relationship exists among the different series. However, in practice one often observes that there exist linear combinations of different unit root time series that behave like stationary series. Examples include: consumption and permanent income in macroeconomic systems, prices and dividends on the stock market, short and long term interest rate series in financial markets. This stylized fact is refered to as cointegration. Since the fundamental work developing the concept of cointegration (e.g. Engle and Granger (1987)), an enormous literature has developed on estimation, testing and prediction in potentially cointegrated models. Most of the work has adopted a classical econometric perspective (see, e.g., Johansen (1995) for an introduction to this literature). However, there has been a substantive amount of Bayesian work on cointegration. The purpose of this paper is to survey and critically assess the Bayesian cointegration literature. The paper is aimed at Bayesians and non-Bayesians. For Bayesians, we offer discussion of the key issues which must be considered when specifying prior distributions and the likelihood and a description of the posterior simulation methods necessary to implement Bayesian methods empirically. As we shall see, the issues the Bayesian must address in cointegration analysis are somewhat different from those addressed by the classical econometrician. Hence, we feel that non-Bayesians should be interested in Bayesian cointegration methods as the different perspective they adopt sheds a new light on many key properties of cointegrated models.

To establish notation and illustrate the basic ideas underlying Bayesian cointegration analysis, let $\{x_t\}_{t=1}^T$ be a realization of the p-dimensional Vector autoregressive (VAR) process of lag length k:

$$x_t = \sum_{i=1}^{k} \Gamma_i x_{t-i} + \Phi d_t + \varepsilon_t, \tag{1}$$

where $\varepsilon_t \overset{iid}{\sim} N_p(0, \Sigma)$.[1] d_t contains deterministic terms (i.e. an intercept, deterministic trends, dummy variables, etc.). This model can be written in error correction model (ECM) form as:

$$\Delta x_t = \Pi x_{t-1} + \sum_{i=1}^{k-1} \Psi_i \Delta x_{t-i} + \Phi d_t + \varepsilon_t, \tag{2}$$

where the matrix of long run multipliers, Π, can be written as $\Pi = \alpha\beta'$ where α and β are both full rank $p \times r$ matrices where $0 \leq r \leq p$ is the number

[1] In this paper, we will assume errors are Normally distributed. Extensions to more flexible distributions (e.g. the Student-t) can easily be obtained by using mixtures of Normal distributions in the standard Bayesian manner. See, for instance, Geweke (1993) or Escobar and West (1995) for examples of such an approach.

of cointegrating relationships. Note that, if $r = p$ then all the elements of x_t are trend-stationary, while if $r = 0$ then the series all contain a unit root, but cointegration does not occur.

In one sense, Bayesian analysis of cointegration is straightforward. Equations (1) or (2) and the accompanying assumptions define the likelihood function. The researcher can combine the likelihood function with a prior and do Bayesian inference with the resulting posterior. However, interesting and empirically important issues of identification (and, as a result, prior elicitation) arise from the fact that Π is potentially of reduced rank. A global identification issue can be seen by noting that $\Pi = \alpha\beta'$ and $\Pi = \alpha A A^{-1}\beta'$ are identical for any nonsingular A. This indeterminacy is commonly-surmounted by imposing the so-called *linear normalization* where $\beta = [I_r \quad B']'$. However, as we shall see, there are some drawbacks to this linear normalization. Even if global identification is imposed, a local identification issue occurs at the point $\alpha = 0$ (i.e. at this point β does not enter the model). As we shall see, this local identification problem can cause serious problems for Bayesian inference. For instance, a common noninformative prior can lead to a posterior distribution which is improper (i.e. is not a valid p.d.f.[2]) thus precluding valid statistical inference. This issue was brought forward by Kleibergen and Van Dijk (1994, 1998). The development of the Bayesian cointegration literature reflects an increasing awareness of these issues and this paper is organized to reflect this development. In particular, we begin by discussing early work, based on VAR or VMA representations which ignored these issues. We then proceed to a discussion of work based on the ECM representation, beginning with a simple specification using the linear normalization and Normal priors before moving onto the recent literature which develops methods for sensible treatment of the identification issues.

2 Early Work: Unrestricted VARs and inference conditional upon the cointegration rank

Much of the earliest Bayesian work on cointegration (e.g. DeJong (1992), Dorfman (1994), Koop (1991, 1994)) did not work with the cointegrated ECM. This simplified Bayesian computation, but meant that the priors used by the authors did not reflect or consider the reduced rank restrictions implied by cointegration. As a representative example of this work, consider DeJong (1992) which worked with a VAR using a noninformative prior for the model parameters. Bayesian VAR methods are simple and well-established and, hence, DeJong (1992) remained within a familiar framework (see, e.g., Zellner (1971)). He used Monte Carlo integration to take random draws from the posterior of $(\Gamma_1, .., \Gamma_k)$ which are then transformed to build up posteriors of the roots of the VAR representation. Cointegration is related to the number of nonstationary roots in the VAR, the probability of which is calculated using output from the Monte Carlo integration procedure. Given future developments, it is interesting to note that

[2] Improper densities are those which do not integrate to one.

DeJong (1992) discusses the properties of his prior and points out that, despite being Uniform and "noninformative" over the VAR coefficients, it is far from Uniform over the VAR roots which are the basis for cointegration inference. The issue that a flat prior in one representation of a model may be very informative in an undesired way in another representation will come back in the next section. It is, however, important to note that in the case where data information on the number of stationary relations in a VAR is so strong that this information follows unambiguously from the posterior of the roots of the VAR, then it follows that the choice of a Uniform prior, restricted by some inequality conditions to the regions where the data are informative, is a sensible choice. In many empirically relevant economic models like the term structure of interest rates and present value models for stock and bond prices, the data information is usually not that informative and more informative priors than the Uniform (over the VAR coefficients) are necessary. Thus an important research area became the search for a weakly informative prior where the information in the likelihood dominates strongly but where the posterior density of parameters of interest and posterior probabilities of model characteristics are well defined.

Influential early work using the ECM includes Bauwens and Lubrano (1996), Geweke (1996) and Kleibergen and van Dijk (1994). Since this work forms the basis of much of the future Bayesian cointegration work, it is instructive to consider these papers in more detail. The key innovation of these papers was to condition on a given number of cointegrating vectors, r, and directly work with α and β. In other words, unlike previous approaches, the reduced rank nature of Π is directly imposed in a posterior simulation algorithm. This allows for Bayesian estimation and inference on α and β (and all other model parameters) for a given r. By carrying out posterior inference for every possible r, the researcher can then use standard Bayesian model comparison methods (e.g. posterior or predictive odds ratios) to select r.

Relative to the VAR, Bayesian inference in the ECM is complicated by the fact that $\Pi = \alpha\beta'$ involves a product of parameters. This precludes direct use of analytical or Monte Carlo integration results for the VAR. However, Bauwens and Lubrano (1996), Geweke (1996) and Kleibergen and van Dijk (1994) note that once we condition on the cointegrating vectors, β, the otherwise nonlinear ECM becomes a linear one. This means that, under suitable informative priors (e.g. Normal priors of the form presented in Geweke (1996)), standard Bayesian analysis of the VAR model applies (conditional on β). Furthermore, for particular specifications, the posterior distribution of β conditional on α,Ψ, Φ and Σ has a standard distribution (where $\Psi = (\Psi_1, .., \Psi_{k-1})$). This suggests that a posterior simulation method known as Gibbs sampling may be set up.

Gibbs sampling starts out from an initial value for all model parameters and then produces a sequence of random draws by cycling through the *full conditional posterior distributions*, always conditioning on the most recent draws of the conditioning parameters. Under weak conditions (verified by Geweke (1996) for the prior he considers) it can be shown that this sequence of draws from the full conditionals converges to a sequence of draws from the joint posterior density. Thus, Gibbs sampler output can be treated in the same fashion as output

from a Monte Carlo integration procedure (e.g. averages of draws converge to posterior means).

A Gibbs sampling algorithm for the cointegrated ECM consists of the following steps.

1. Initialize all parameter matrices: $\alpha^{(0)}, \beta^{(0)}, \Psi^{(0)}, \Phi^{(0)}$ and $\Sigma^{(0)}$ with, for example, the maximum likelihood estimates in Johansen (1995).

2. Generate a draw $\Sigma^{(1)}$ from the posterior distribution of Σ conditional on $\alpha^{(0)}, \beta^{(0)}, \Psi^{(0)}, \Phi^{(0)}$.

3. Generate a draw $(\Psi^{(1)}, \Phi^{(1)})$ from the joint posterior distribution of Ψ and Φ conditional on $\alpha^{(0)}, \beta^{(0)}$ and $\Sigma^{(1)}$.

4. Generate a draw $\alpha^{(1)}$ from the joint posterior distribution of α conditional on $\beta^{(0)}, \Psi^{(1)}, \Phi^{(1)}$ and $\Sigma^{(1)}$.

5. Generate a draw $\beta^{(1)}$ from the joint posterior distribution of β conditional on $\alpha^{(1)}, \Psi^{(1)}, \Phi^{(1)}$ and $\Sigma^{(1)}$.

6. Repeat steps 2-5 until the sequence of draws $\{\alpha^{(i)}, \beta^{(i)}, \Psi^{(i)}, \Phi^{(i)}, \Sigma^{(i)}\}_{i=1}^{N}$ is large enough to provide an accurate approximation of the properties of $p(\alpha, \beta, \Psi, \Phi, \Sigma | Data)$.

Geweke (1996) also discusses methods of evaluating the accuracy of estimates of posterior properties of functions of the parameters (e.g. posterior means) produced using the Gibbs sampler. The posterior density itself may be estimated from $\{\alpha^{(i)}, \beta^{(i)}, \Psi^{(i)}, \Phi^{(i)}, \Sigma^{(i)}\}_{i=1}^{N}$ using simple histograms or more sophisticated density estimates. Posterior moments are, in most cases, easily computed by arithmetic averages.

Geweke (1996) addressed the global identification issue described in Section 1 through linear normalizations such as $\beta = [I_r \quad B']'$. He showed that, if standard informative priors are used, the Gibbs sampling algorithm is of a particularly simple form involving only draws from the multivariate Normal distribution (in Steps 3, 4 and 5) and the inverted Wishart (in Step 2). The standard prior considered by Geweke (1996) involved Normal forms for α, B, Ψ and Φ and an inverted Wishart form for Σ.

Even if the linear normalization is retained, the sampling scheme outlined above is not generally applicable for all of the priors for B that one may wish to choose. However, given the standard form of the posterior for $(\alpha, \Psi, \Phi, \Sigma | B)$, once a draw of B is obtained, the drawing of the remaining parameters is straightforward regardless of the form of the prior for B. This suggests, then, that a method of sampling B from its marginal posterior distribution is required. To this end, Bauwens and Lubrano (1996) present an approach to obtaining draws of B using importance sampling, while Bauwens and Giot (1998) demonstrate that a particular sort of Gibbs sampler known as griddy-Gibbs works well.

Bauwens and Lubrano (1996) and Geweke (1996), which developed computational methods which allowed for simple and efficient Bayesian inference in the cointegrated ECM under a range of commonly-used priors (including an apparently "noninformative" one), seemed to complete the research project establishing the basic tools for Bayesian analysis of cointegrated models. However, it soon became clear that there were problems with these early approaches (some of which were noted already by Kleibergen and van Dijk (1994)), which stimulated a flurry of additional approaches. Before describing these approaches, we first explain what are these problems.

3 Problems with the early work: Identification and normalization issues

As we have seen in Section 2, several standard prior distributions have emerged in the early Bayesian cointegration literature. Most of them are either noninformative or *conjugate*, *i.e.* priors with the attractive property of leading to posterior distributions in the same parametric family as the prior. Even when non-conjugate informative priors are used, it is common to stay within familiar frameworks which make computation easy. Typically, in Normal likelihood models, the use of Normal priors over regression coefficients and inverted-Wishart priors over error covariance matrices is particularly convenient. These prior choices are made, e.g., in Geweke (1996). When designing priors for a new model class, it is common practice to use such standardized priors, at least as a first attempt, and the cointegrated VAR is no exception.

There are two main features of the cointegrated ECM which make the standard priors unsuitable for cointegration analysis, however. First, the reduced rank restriction of the cointegrated ECM introduces a rather complex nonlinearity in the otherwise linear VAR. Most standard priors have been developed and evaluated on linear models and there is no guarantee that their properties will carry over to nonlinear models. Second, the cointegrated ECM is inherently non-identified in the sense that the cointegration vectors are only uniquely determined up to arbitrary linear combinations (see Section 1), i.e. data only carries information about the cointegration space. This means that the cointegration vectors must be restricted for identification and that the set of unrestricted elements of β that remain to be estimated depends on the chosen identification scheme. As mentioned, a common choice of identifying restrictions is the set of linear restrictions. To implement these restrictions, we assume we know which r rows of β will be linearly independent, partition $\beta = \left(\beta_1', \beta_2'\right)'$ where β_1 is $r \times r$ and specify a selection matrix c such that $c\beta$ is invertible ($|c\beta| \neq 0$). Next, we normalize upon these rows by $\beta\left(c\beta\right)^{-1}$. For example, assume we select the first r rows ($c = [I_r\, 0]$), then the resulting specification for β will be the linear normalization $\beta = [I_r\ B']'$ where $B = \beta_2\beta_1^{-1}$. This, however, will have implications for the specification of α as these two parameters always occur in the model as the product $\alpha\beta'$. The choice of prior is thus intertwined with the choice of

identification, which makes the prior specification a rather delicate problem.

An example of how a straightforward adoption of traditional priors to the cointegrated ECM brings unwanted distortion of prior beliefs is given by Strachan and van Dijk (2004). They show that a flat and apparently "noninformative" prior on B in the linear normalization $\beta = [I_r \; B']'$ favors the cointegration spaces near the region where the linear normalization is invalid ($|c\beta| = 0$). Hence, the linear normalization is used under the assumption that it is valid while at the same time the prior says that the normalization is very likely to be invalid.

Assuming we use the linear identifying restrictions, another important issue identified by Kleibergen and van Dijk (1994) is that of local non-identification. The issue here is that when α has reduced rank (e.g., $\alpha = 0$) the conditional posterior distribution for $B|\alpha$ is equal to its prior (i.e. since B does not enter the likelihood function at the point $\alpha = 0$ there is no data-based learning about B and, thus, its posterior equals its prior at this point). If the prior for $B|\alpha = 0$ is improper (as it is in the common "noninformative" case), then the posterior will also be improper. Formally, Kleibergen and van Dijk (1994) associate the local non-identification problem with nonexistence of posterior moments and non-integrability of the posterior (under a common noninformative prior). Kleibergen and van Dijk (1998) additionally point out that local non-identification implies an absorbing state in a Gibbs sampler, thereby violating the convergence conditions for the sampler.

Another issue with the identifying restrictions is the validity of the chosen normalization. Put differently, when we select c (and so the rows of β upon which to normalize), we may make an invalid selection such that $c\beta$ is singular. There has been much work in the classical framework addressing this issue (see, for example, Boswijk (1996) and Luukkonen, Ripatti and Saikkonen (1999)), but a Bayesian investigation is provided in Strachan (2003). Finally, even if the chosen normalization is valid, the class of models that may be considered is restricted to exclude some that have proven very important in cointegration analysis generally. For example, Strachan and van Dijk (2004) show that imposing weak exogeneity results in an improper posterior when an uninformative prior is used, but also when particular informative priors are used.

4 Prior distributions for the cointegrated ECM

As demonstrated in the previous section, priors placed upon the elements of the cointegrating vectors B, even ones which appeared to be "noninformative", have many important (and often undesirable) impacts upon empirical analysis. In this section, we describe three key recent approaches for surmounting many of the problems outlined in the previous section. The first, which we only briefly discuss, is the Jeffreys' prior approach (e.g. Kleibergen and van Dijk (1994)). The second is the embedding approach due to Kleibergen (1997), Kleibergen and van Dijk (1998), Kleibergen and Paap (2002), and Kleibergen (2004)) and extended in Strachan (2003). These approaches deal variously with issues of

local nonidentification and the possibility of using invalid identifying restrictions. The third line of research, see Villani (2000) and further developed in Villani (2004), Strachan (2003), Strachan and Inder (2004) and Strachan and van Dijk (2004), shifts the focus from the cointegrating vectors to treating the cointegrating space sp(β) as the object of interest.

4.1 The Jeffreys' prior approach

In proposing the use of the Jeffreys' prior, Kleibergen and van Dijk (1994) were motivated by the result that at points of local nonidentification (such as where α has reduced rank) the Jeffreys' prior is zero. Thus, these problematic points are excluded from the support of (α, B) in the posterior distribution. This approach has been extended by Martin (2001) to a fractional cointegration model. The Jeffreys' prior has the additional advantage that it is invariant with respect to certain specifications of the model and thus is more attractive than a Uniform prior. The posterior of several parameters of interest (such as r and k in (1)) becomes improper, however, due to the Jeffreys' prior being improper (see Subsection 5.3 for further discussion on this issue commonly termed Bartlett's paradox). In light of this, posterior probabilities on sharp nulls are not well defined (e.g. selecting a value for r cannot be done using posterior odds ratios). This point, which lessens the usefulness of the Jeffreys' prior for cointegration analysis, will be elaborated on in Section 5. Another problem when using the Jeffreys' prior is treatment of initial conditions. This matter is discussed in Kleibergen and van Dijk (1994) where a number of possible solutions are outlined.

4.2 The Embedding Approach

The embedding approach is a way of addressing the problems arising from the local identification problem discussed in Section 3. This approach is based upon the insight that while Π is globally identified, the problems arising from local non-identification arise since Π has reduced rank if cointegration occurs. However, a so-called *embedding model* can be constructed which nests the ECMs for various values of r and, crucially, contains a matrix of parameters, λ, which reflects the degree of rank reduction. In this approach, prior elicitation is carried out by first eliciting a prior for the unrestricted embedding model, *i.e.*, $p(\Pi)$. Then a transformation from Π to α, β and λ is carried out to obtain the prior $p(\alpha, \beta, \lambda)$. Priors for the reduced rank ECMs are then obtained as this prior conditional on $\lambda = 0$, $p(\alpha, \beta | \lambda = 0)$. As we shall see in this subsection, such an approach has many desirable properties.

Kleibergen and van Dijk (1994) investigate the implications of employing a simple variable addition specification for the restriction of the $p \times p$ long run multiplier matrix Π to reduced rank. They consider the decomposition

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \begin{bmatrix} I_r & B' \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}$$

where Π_{11} and α_1 are $r \times r$ matrices, Π_{12}, Π_{21}' and α_2 are $r \times (p-r)$ matrices and Π_{22} and λ are $(p-r) \times (p-r)$ matrices. The embedding model is (2) with this specification for Π . Note that, if $\lambda = 0$, then this model is the cointegrated ECM with r cointegrating vectors. The embedding approach involves putting a prior over the parameters in the embedding model, and the prior for the ECM with r cointegrating vectors is derived from this prior conditional on $\lambda = 0$.

The validity of this approach relies upon the following argument. If the matrix Π has reduced rank $r < p$, then it has r linearly independent rows and columns. If we know which rows and columns these are, we may rearrange the matrix such that these are the first r rows and columns and Π_{11} is of full rank r and we have the definitions $\alpha_1 = \Pi_{11}$, $\alpha_2 = \Pi_{21}$, $B' = \Pi_{11}^{-1}\Pi_{12}$ and $\lambda = \Pi_{22} - \Pi_{21}\Pi_{11}^{-1}\Pi_{12}$. In this case, if Π has reduced rank r (and, thus, $\lambda = 0$) then this implies[3] $\Pi_{22} = \Pi_{21}\Pi_{11}^{-1}\Pi_{12}$.

A local identification problem occurs where $\alpha_1 = \Pi_{11} = 0$ and, as we have seen in Section 3, this causes problems for Bayesian analysis. If we knew which rows and columns of Π are linearly independent we could simply exclude from the support of Π_{11} the point where $|\Pi_{11}| = 0$. Problems arise, however, when we do not know which rows and columns of Π are linearly independent - which occurs in most practical situations. Kleibergen and Paap (2002) also link this approach to the problem of non-uniqueness of the posterior distribution of α and B conditional upon rank reduction, a manifestation of the Borel-Kolmogorov paradox.

For this reason, a direct application of the embedding approach along the lines outlined above is problematic. However, Kleibergen and van Dijk (1998) and subsequently Kleibergen and Paap (2002) address this issue by proposing a singular value decomposition. To see how this avoids the problem of not knowing which rows and columns of Π are linearly independent, consider the singular value decomposition of Π as

$$\Pi = USV' = U_1 S_1 V_1' + U_2 S_2 V_2'$$

where $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$, $U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$, $S = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}$ and S is a diagonal matrix with the (positive) singular values of Π in descending order down the diagonal and both U and V are orthonormal matrices such that $U'U = V'V = I_p$. Further, U_{11}, S_1 and V_{11} are $r \times r$ matrices such that given Π is a $p \times p$ matrix, the dimensions of the remaining matrices are subsequently defined.

>From the singular value decomposition above, the elements of the cointegrating model are defined by Kleibergen and Paap (2002) as $\alpha' = V_{11} S_1 \begin{pmatrix} U_{11}' & U_{21}' \end{pmatrix}$, $B = V_{11}^{-1} V_{12}$ and $\lambda = (V_{22}' V_{22})^{-1/2} V_{22} S_2 U_{22}' (U_{22} U_{22}')^{-1/2}$. The decomposition of Π then becomes

$$\Pi = \alpha\beta' + \alpha_\perp \lambda \beta_\perp'$$

[3] This result must hold as at rank r with Π_{11} full rank, then each ij^{th} element of $\Pi_{22} = [\pi_{22,ij}]$ is determined by the determinant $\begin{vmatrix} \Pi_{11} & \pi_{12,j} \\ \pi_{21,i} & \pi_{22,ij} \end{vmatrix} = 0$ where $\pi_{12,j}$ is the jth column of Π_{12} and $\pi_{21,i}$ is the ith row of Π_{21}.

such that a rank of r implies $S_2 = 0$ and so $\lambda = 0$ and $\Pi = \alpha\beta'$. The implementation of this approach requires specification of the posterior for the full rank Π, $p_\pi(\Pi|y)$, and then the transformation to (α, B, λ) to obtain the posterior

$$p_\theta(\alpha, B, \lambda|y) = p_\pi\left(\alpha\beta' + \alpha_\perp\lambda\beta'_\perp|y\right)|J(\Pi : \alpha, B, \lambda)|,$$

where $|J(\Pi : \alpha, B, \lambda)|$ is the Jacobian for the transformation from Π to (α, B, λ). The reduced rank model obtains by considering this distribution at the point $\lambda = 0$.

It is interesting to note the relationship between the Jeffreys' prior and embedding approaches. As we have seen, an attractive property of the Jeffreys' prior is that it is zero at points of local nonidentification (such as where α has reduced rank) and, thus, these points are excluded from the support of (α, B) in the posterior distribution. The singular value approach uses a similar (related) behavior of the Jacobian for the transformation from Π to exclude points of local nonidentification from the support of the posterior. That is, when α has reduced rank, $|J(\Pi : \alpha, B, \lambda)| = 0$. Kleibergen and Zivot (2003) show that the Jeffreys prior results from the embedding approach when the embedding specification is imposed upon $\Pi^* = \left(\Sigma x_{t-1}x'_{t-1}\right)^{\frac{1}{2}}\Pi\Sigma^{-\frac{1}{2}}$ and a flat prior is specified on Π^*.

The embedding approach was a significant advance in Bayesian cointegration analysis in overcoming the problems associated with local non-identification and results in a prior that is invariant to the chosen normalisation. However, the linear normalization was used in its development and, thus, does not address the problem of global non-identification. This point is made in detail in Strachan (2003), which links the use of linear identifying restrictions to a range of the problems (some discussed in Section 3), and in particular to the issue that using linear identifying restrictions places a restriction on the estimable region of the cointegrating space (see Subsection 5.3 for further discussion). Strachan and Inder (2004) provide an extensive discussion of further problems associated with the use of linear identifying restrictions. Strachan (2003) therefore proposes a specification of the identifying restrictions in an embedding approach that does not restrict the estimable space and so can estimate any element of the Grassman manifold and implies a Uniform distribution over the Grassman manifold. The work discussed in the following section takes this concept further.

4.3 The Grassman Approach: Placing a Prior on the Cointegrating Space

The global identification problem (i.e. that there are an infinite number of ways of carrying out the decomposition $\Pi = \alpha\beta'$) mentioned in the introduction implies that the cointegration vectors are only identified up to arbitrary linear combinations. Thus, only the space spanned by the cointegrating vectors, $\mathfrak{p} = sp(\beta)$, can be estimated from data. Taking the cointegration space as the fundamental entity in cointegration models leads naturally the view expounded in Villani (2000) that a prior on the cointegration vectors should be evaluated by how it distributes probability mass across the support of $\mathfrak{p}$. For example, it is

natural to use a Uniform distribution over the support of $\mathfrak{p}$ to express ignorance. However, as demonstrated in Strachan and Inder (2004), a Uniform prior on B implies a very different informative, and undesirable, prior distribution on $\mathfrak{p}$. Developing a prior for $\mathfrak{p}$ requires, therefore, a way of visualizing the parameter $\mathfrak{p}$, its support, and ways of placing priors upon this support.

Formally and generally, the cointegrating space $\mathfrak{p}$ is an r-dimensional hyperplane in an p-dimensional space and its relation to the cointegrating vectors β is that these vectors lie in and thereby identify that plane. To assist in visualizing the parameter $\mathfrak{p}$, we give two simple examples. Consider first the case where $p = 2$ and $r = 1$ such β is a 2×1 vector. This vector is shown in Figure 1 as the black line with the open arrow head. The space spanned by this vector $\mathfrak{p} = sp(\beta)$ is the dashed line in which the vector lies. The support of $\mathfrak{p}$ is, in this case, the collection of all possible lines passing through the origin. To generalize this slightly, consider the case $p = 3$ and $r = 2$ such that β is a 3×2 matrix and each of the 3×1 vectors are plotted in Figure 2. In this case $\mathfrak{p}$ is the 2-dimensional cross-hatched plane in which the vectors lie and the support of $\mathfrak{p}$ is all of the 2-dimensional planes passing through the origin such that all directions for the planes are covered.

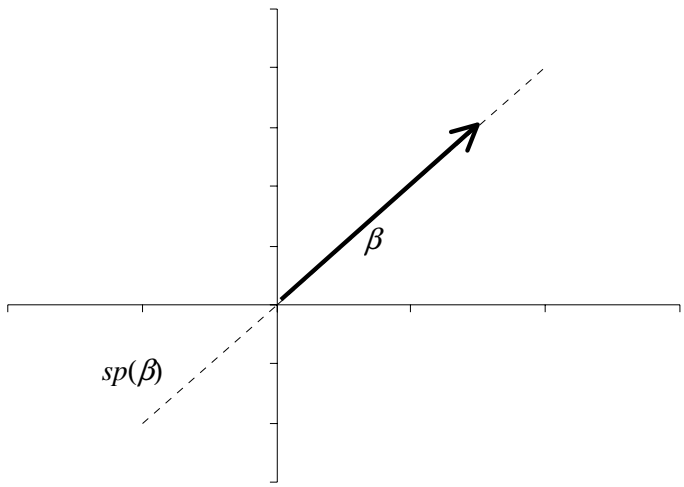

Figure 1: The cointegrating space is the one-dimensional plane, or line, represented by the dashed line. The vector lying in the cointegrating space is a cointegrating vector.

The support for $\mathfrak{p}$ (which we will formally define in a moment) in general, then, is a rather abstract space. Three questions come to mind: i) is there a non-controversial definition of a Uniform distribution over the set of cointegration spaces, ii) does such a distribution exist, and iii) is it unique? All three of these questions can be answered in the affirmative and the appropriate mathematical

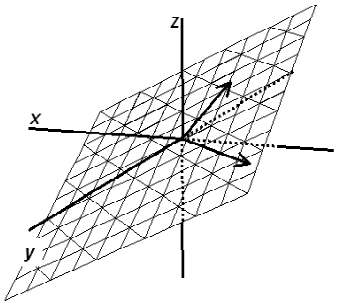

Figure 2: Two dimensional cointegrating space represented by the cross-hatched plane. The two vectors lying in the plane are therefore cointegrating vectors.

apparatus has been around for nearly a century.

In order to describe this line of research we revert to the simplest case in Figure 1 of a bivariate process with a single cointegration vector parameterized in polar coordinates $\beta = (\cos\theta \ \sin\theta)'$,where $\theta \in [-\pi/2, \pi/2)$. The length of β is restricted to unity for identification. As shown in Strachan and Inder (2004), this identifying restriction does not distort the weight on the space of the parameter of interest, $\mathfrak{p}$. A natural candidate for a noninformative distribution on $\mathfrak{p}$ is the Uniform distribution on θ, the parameter governing the direction of β and therefore $\mathfrak{p}$.

The next step is to extend the concept of Uniform distribution to the general case of an arbitrary number of cointegrating relations. In this case a Uniform distribution for $\mathfrak{p}$ will no longer be implied by a Uniform distribution on angles in higher dimensions. Our aim is thus to arrive at a rigorous definition of the intuitive idea of assigning equal prior probability to every possible cointegration space of dimension r. In order to do this we need to introduce a few concepts. Our treatment here will focus on ideas and we refer to James (1954) for background details.

The set of all $p \times r$ orthonormal matrices is called the *Stiefel manifold* $\mathbb{V}_{p,r}$. Two special cases of the Stiefel manifold are the unit sphere ($r = 1$) and the *orthonormal group* $\mathbb{O}_p$, the set of $p \times p$ orthonormal matrices ($r = p$). The Stiefel manifold is a compact space and admits a Uniform distribution. In the case where $r = 1$, one might conceptualise the collection of directions of all p-dimensional unit vectors, $\beta \in \mathbb{V}_{p,1}$, as describing a p-dimensional unit sphere centered at the origin. Thus, we may visualise a Uniform distribution on the p-dimensional unit sphere as characterising a Uniform distribution on $\mathbb{V}_{p,1}$. For $r > 1$, we can think of each vector in β as describing a unit sphere with the additional restriction that the vectors are all orthogonal to each other.

The *Grassman manifold* $\mathbb{G}_{p,r}$ is the abstract space of all possible r-dimensional

planes of $\mathbb{R}^p$. An example of an element of $\mathbb{G}_{3,2}$ is given as the two-dimensional cross-hatched plane in Figure 2. This defines the cointegrating space (the parameter of interest) as an element of the Grassman manifold (the support of the parameter), that is $\mathfrak{p} \in \mathbb{G}_{p,r}$. As stated earlier, in the ECM only the space spanned by the columns of β is identified, such that we only have information on $\mathfrak{p} = sp(\beta) \in \mathbb{G}_{p,r}$. A Uniform prior for the cointegration space is therefore given by the Uniform distribution on $\mathbb{G}_{p,r}$.

The next step is to make the notion of a Uniform prior distribution on $\mathbb{G}_{p,r}$ operational for a Bayesian analysis. Villani (2000, 2004) develops the Uniform prior on the cointegrating space in the linear normalization of β. He shows that the Uniform distribution on the Grassman manifold implies a matrix Cauchy distribution on B. Thus a Uniform distribution on $\mathbb{G}_{p,r}$ implies $p(B) \propto |I_r + B'B|^{-p/2}$.

The linear normalization may be regarded as convenient in many respects, but, as discussed in Strachan and Inder (2004) and Strachan and van Dijk (2004), has a number of drawbacks (some discussed in Section 3, but also some others which we will not elaborate on here). These papers develop an alternative approach which avoids the use of linear identifying restrictions through a more direct method of eliciting a prior on $\mathbb{G}_{p,r}$. This approach uses the natural relationship between the Grassman manifold and the Stiefel manifold and the development of measures on these spaces presented in James (1954). In particular, they use the result that the Uniform distribution on $\mathbb{V}_{p,r}$ induces a Uniform distribution on $\mathbb{G}_{p,r}$ (James (1954) and Strachan and Inder (2004)). Thus, it is possible to work with semiorthogonal matrices, *i.e.* $\beta \in \mathbb{V}_{p,r}$, and adjust all integrals by dividing with the volume of $\mathbb{O}_r$ (James (1954) and Strachan and van Dijk (2004)) to account for the fact that $\mathbb{V}_{p,r}$ is a larger space than $\mathbb{G}_{p,r}$. A convenient feature of this approach is no normalisation is imposed, but if a normalisation is desired, this may be imposed after estimation of β.

So far, all our discussion in this subsection focusses on the prior for the cointegration space and ignores the other parameters in the model. For completeness, we should mention that the prior distribution on Σ, Γ and Φ may be chosen in many ways. For instance, ideas from the traditional VAR literature may be used (e.g. the well known Minnesota prior in Litterman (1986) can be used). A common reference prior for the adjustment coefficients is $\alpha_i|\Sigma \overset{iid}{\sim} N(0, \sigma_\alpha^2 \Sigma)$, where α_i contains the adjustment coefficients for the ith cointegrating relation in the semiorthogonal normalization of β (see Strachan and Inder (2004) and Villani (2004)). The hyperparameter σ_α controls the shrinkage toward the zero matrix.

4.3.1 Informative priors on the cointegrating space.

Although the Uniform prior is often used in theoretical work, it is common for practitioners to employ informative priors for parameters. Strachan and Inder (2004) therefore present a method of eliciting an informative prior for $\mathfrak{p}$. The objectives of the method they propose are to place the mass of the

prior distribution upon some preferred location, and to specify some level of dispersion for this distribution. The discussion here is cursory for brevity and we refer the reader to Strachan and Inder (2004) for details.

If a researcher believes a parameter is likely to have a particular value, to incorporate this prior belief she places more prior mass around this likely point. Here the "parameter" is the cointegrating space, $\mathfrak{p}$, and it will often be the case that a researcher has prior beliefs about plausible regions of this space. As a simple example, consider a vector of four interest rates $x_t = (x_{1,t}\ x_{2,t}\ x_{3,t}\ x_{4,t})$. Various expectations theories of the term structure of interest rates imply that if each of the $x_{i,t}$ has a unit root, the spreads, $x_{i,t} - x_{j,t}$, will be stationary. This would suggest that such prior beliefs can be expressed through a matrix H (see Johansen (1995) for examples and Strachan and van Dijk (2003) for an application) such as

$$H = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Since $sp(H) = sp(H\kappa)$ for any full rank square κ, after constructing H, we may innocuously map H to $V_{r,n}$ by the transformation $H \to H(H'H)^{-1/2}$. For the parameter $\mathfrak{p}$ then, denote the likely or location value as $\mathfrak{p}^h = sp(H)$.

A dogmatic prior for $\mathfrak{p}$ could be obtained by letting $\beta = H$ and this prior assigns probability one to the point $\mathfrak{p} = \mathfrak{p}^h$. However it is common that the researcher will want to employ a less dogmatic prior. In our example, we may wish the random space $\mathfrak{p}$ to have a mean under the prior of $\mathfrak{p}^h$, but be allowed to vary over the entire Grassman manifold. This is achieved by introducing a random scalar τ which performs a number of roles. First, it acts as a weight such that mass of the prior distribution is distributed over the Grassman manifold between the space of H and the space of $H_\perp$ which lies in the orthogonal complement of $\mathfrak{p}^h$. Second, the variance of τ controls the dispersion of the distribution of $\mathfrak{p}$ around the location $\mathfrak{p}^h$. Finally, as the sign of τ may be positive or negative with equal probability, probability mass is allocated to all regions of the Grassman manifold, including between the spaces of H and $-H_\perp$.

One method of developing an informative prior follows. Let the random scalar τ have $E(\tau) = 0$ and $E(\tau^2) = \sigma_\tau^2$. For example, we may choose $\tau \sim N(0, \sigma_\tau^2)$. The value of σ_τ will control the tightness of the prior around $\mathfrak{p}^h$. Next construct $P_\tau = HH' + H_\perp H_\perp' \tau$ and let the $n \times r$ matrix Z be distributed as $vec(Z) \sim N(0, I_{nr})$. The matrix $X = P_\tau Z$ can be decomposed as $X = \beta\kappa$ where $\beta \in V_{r,n}$ and κ is an $r \times r$ lower triangular matrix. The resulting distribution for $\mathfrak{p} = sp(\beta)$ will then be centred upon $\mathfrak{p}^h$ with its dispersion around this point determined by the value chosen for σ_τ. The reader interested in more detail is referred to Strachan and Inder (2004).

This completes our discussion of prior elicitation. Of course, a researcher may choose any prior she wishes to reflect her prior beliefs. In this section, we have described several common choices, ending with a discussion of prior elicitation (both informative and noninformative) directly over the cointegration

space. We argue that this last approach, which surmounts the problems caused by global and local non-identification in a sensible manner, will be the preferable one to use in most applications. Now that we have interesting classes of priors and a likelihood (given by equation 2 and the assumption of Normal errors), we can now proceed to posterior analysis.

5 Posterior distributions: $p(r|y)$ and $p(\beta|r,y)$.

In Bayesian cointegration analysis there are usually two primary aims. The first is to determine the dimension of the cointegrating space, r, and the second is to obtain an estimate of that space. From the decomposition $p(\beta, r|Data) = p(\beta|Data, r)\, p(r|Data)$, we can frame our discussion in terms of $p(\beta|Data, r)$ and $p(r|Data)$. In this section, we discuss both analytical expressions (where they exist) and posterior simulation algorithms. The final part of this section discusses point estimation of the cointegrating space. This is a non-trivial issue in cointegration analysis due to the non-identifiability of the cointegration vectors and, as we shall see, methods that might suggest themselves intuitively are, in fact, inappropriate.

5.1 Analytical results

As we have seen in Section 2, most Bayesian treatments of the ECM use posterior simulation methods such as Gibbs sampling or importance sampling. However, a few analytical results have been derived and it is constructive to examine them to help understand the properties of the posterior in cointegrated models.

Bauwens and Lubrano (1996) show that the marginal posterior of β under a common noninformative prior $p(\alpha, \beta, \Gamma, \Phi, \Sigma) \propto |\Sigma|^{-(p+1)/2}$ is of the form

$$p(\beta|Data, r) \propto \frac{\left|\beta' C_1 \beta\right|^{l_1}}{\left|\beta' C_2 \beta\right|^{l_2}}, \tag{3}$$

where C_1 and C_2 are $p \times p$ data matrices and l_1 and l_2 are scalars (see Bauwens and Lubrano (1996) for precise definitions). In terms of B in the linear normalization $\beta = (I_r, B')'$, this density can be written as ratio of two matrix t density kernels. This type of density has been termed a 1-1 poly matrix t density by Drèze (1978). Villani (2004) shows that, for the specification used in that paper, the marginal posterior of B with a Uniform prior on the Grassman manifold (see Section 4.2) and a proper prior on α, is of the same form with different parameters C_1, C_2, l_1 and l_2.

The posterior distribution of B in (3) is integrable, but lacks moments of any order (see Kleibergen and van Dijk (1994) or Bauwens and Lubrano (1996)). Villani (2004) points out this is a natural feature of the prior in the linear normalization. To see this, let $\beta \in \mathbb{V}_{p,r}$ such that a Uniform distribution on $\mathbb{V}_{p,r}$ implies a Uniform distribution on $\mathbb{G}_{p,r}$. Then B equals the matrix quotient $\beta_2 \beta_1^{-1}$, where β_1 and β_2 are the upper $r \times r$ and lower $(p-r) \times r$ submatrices

of β, respectively, with a Cauchy form for the Jacobian for this transformation $\beta \to B$, and B will therefore have very fat (Cauchy-like) tails in the posterior. In explaining why the maximum likelihood estimator occassionally produces very unreliable estimates of B, Phillips (1994) makes the same point in proving that this estimator has Cauchy tails. By contrast, Strachan and van Dijk (2004) prove that all posterior moments of β are finite in their approach with β specified as a semiorthogonal matrix, $\beta \in \mathbb{V}_{p,r}$.

The analytical posterior results for B are useful for establishing some vital properties of the posterior, such as integrability, but are of limited value to practitioners as the class of poly matrix t distributions is, to a large extent, unexplored beyond the dimension $r = 1$. For example, marginal distributions and moments of poly matrix t variates (the case where $r > 1$) are not known. With the semiorthogonal specification for β of Strachan and Inder (2004) many more analytical results can be obtained. However, analytical results for $p(\beta|Data, r)$ are not enough to deliver posterior distributions of various features of interest, such as forecasts and impulse response functions, which are an important components in applied work. For these reasons, most Bayesian researchers use posterior simulation.

5.2 Posterior Simulation in the Cointegrated ECM

The Gibbs and importance samplers discussed previously (e.g. those in Geweke (1996) and Bauwens and Lubrano (1996) described at the end of Section 2) were designed for models with computationally convenient priors. However, with the more sophisticated (and reasonable) priors of Section 4, they cannot be directly applied. Kleibergen and van Dijk (1998) point out a complication due to the presence of local non-identification that an inappropriately designed Gibbs sampler will have an absorbing state an so produce a reducible Markov chain, thus violating the conditions necessary for convergence of the chain to the posterior. Accordingly, it is useful to offer a more general discussion of posterior simulation methods. As discussed in Section 2, the key challenges for posterior computation arise due to α and β entering the ECM in a nonlinear fashion. Hence, much of this discussion is focussed on these parameters. To simplify notation, we will collect the parameters other than α, β and r into the vector θ.

Posterior simulation with respect to θ and α is generally straightforward and integration with repect to (θ, α) can often be achieved analytically as the posterior distribution for elements of (θ, α) (conditional upon β and r) belong to well understood standard classes of distributions such as the multivariate Normal, inverted Wishart or Generalized Student-t distributions (see for example Zellner (1971) or Bauwens and van Dijk (1990)). For these parameters, methods such as those outlined in Section 2 can be used. Notable exceptions to this case are the embedding specifications which incorporate a complicated Jacobian term such as Kleibergen and van Dijk (1998), Kleibergen and Paap (2002), and Strachan (2003) (see Subsection 4.2) and the Jeffreys' prior approach of Kleibergen and van Dijk (1994).

In contrast to $p(\theta, \alpha | Data, \beta, r)$, posterior simulation of β is not straightforward. There are three general approaches that have been taken. The first is to draw from $p(\beta | Data, \theta, \alpha, r)$ in the context of an MCMC algorithm. Examples of this approach are Geweke (1996), Kleibergen and van Dijk (1998), Kleibergen and Paap (2002) and Strachan (2003). With the computationally convenient prior of Geweke (1996), $p(\beta | Data, \theta, \alpha, r)$ is Normal and Gibbs sampling is straightforward. However, for the other three papers, $p(\beta | Data, \theta, \alpha, r)$ is not of a simple form form and a standard Gibbs sampler cannot be employed. Thus, in these papers, a Metropolis-Hastings algorithm (see Chib and Greenberg (1995)) is developed in which the candidate density is derived from the full rank VAR.

A second approach uses the standard form for $p(\theta, \alpha | Data, \beta, r)$ to analytically integrate with respect to (θ, α). Then draws from $p(\beta | Data, r)$ are taken. If $r = 1$, such that B has a 1-1 poly-t density, moments may be approximated using the approach detailed in Richard and Tompa (1980). For $r \geq 1$, Bauwens and Lubrano (1996) used an importance sampling scheme (see Kloek and van Dijk (1978) or Geweke (1989)) to obtain draws from the posterior distribution of B. Alternatively, by showing that the conditional densities for $(\beta_1 | \beta_2, \ldots, \beta_r)$, $(\beta_2 | \beta_1, \beta_3, \ldots, \beta_r)$, ... and so on, are 1-1 poly-t, they suggest a Gibbs sampling approach could be used and such an algorithm is presented in Bauwens and Richard (1985). This approach is useful only for a Uniform prior on B, and so Bauwens and Giot (1998) present a griddy-Gibbs sampler for obtaining draws on B for more general priors. Strachan and van Dijk (2004), who use both Uniform and informative priors over the cointegration space, also work with $p(\beta | Data, r)$ but use a Metropolis-Hastings algorithm.

A third approach is presented by Villani (2004) in which θ is analytically integrated out and an MCMC scheme developed to draw from $p(\beta | Data, r, \alpha)$ and $p(\alpha | Data, r, \beta)$. This approach, which uses the linear normalization $\beta = [I_r \; B']'$, takes advantage of the result that the conditional densities for $\alpha | B$ and $B | \alpha$ are Generalized Student t such that a Gibbs sampler may be used.

An important advantage of a simulation based evaluation of the posterior distribution is that the posterior distribution of any quantity which is a function of the parameters and the data can easily be computed from the generated posterior sample. So, for instance, posterior properties of impulse response functions and forecasts are immediately available.

5.3 Posterior distribution of the cointegration rank

It is common to select a value for r (e.g. the posterior mode) and then base empirical results (e.g. about the cointegration space) conditional on this choice. However, one of the advantages of the Bayesian approach is that uncertainty about this parameter may be incorporated into the analysis by using $p(r | Data)$ to average across the various models. These two approaches relate respectively to model selection and Bayesian model averaging, the latter being a topic we will discuss in more detail below. Regardless of whether the researcher adopts a model selection or model averaging approach, it is necessary to know $p(r | Data)$.

In this subsection, we will discuss several ways of learning about this density. Before we do this, the issue of Bartlett's paradox should be mentioned in passing. This issue arises when two models of different dimensions are being compared and is relevant here because deriving $p(r|Data)$ can be interpreted as comparing models of different dimension (e.g. $r = 1$ and $r = 2$ describe ECMs of different dimension). Bartlett's paradox is often informally expressed as saying improper priors[4] should be avoided when calculating Bayes factors (except over parameters common to both models), since then the Bayes factor can depend on arbitrary integrating constants. So, for instance, Bartlett's paradox implies that, with the Jeffreys' prior approach, $p(r|Data)$ cannot be obtained. Some of the technical issues in the recent literature relate to this point, although we will not elaborate on it in any detail in this survey.

Nesting the reduced rank cointegrated ECM within the embedding model (i.e. the unrestricted VAR), Kleibergen and Paap (2002) use MCMC output from a Metropolis-Hastings sampler and the Savage-Dickey density ratio (see Verdinelli and Wasserman (1995)) to estimate the Bayes factors for the model with rank $r < p$ to the model with full rank $r = p$. Using the embedding prior of Subsection 4.2 makes estimation from the reduced rank model and its marginal likelihood computationally easy. However the computation of the Bayes factors does require estimation of a correction factor (which is not defined for the improper prior due to Bartlett's paradox) for the ECM with $r < p$. Examples of the application of this approach are given in Paap and van Dijk (2003) and Strachan (2003). Interestingly, Kleibergen (2004) shows by use of Hausdorff measures and integrals that the posterior odds ratio developed in this way is well defined even with improper priors on the parameters of the full rank model (thus avoiding Bartlett's paradox) and negates the need to estimate the correction factor for the improper prior. The understanding of this argument requires that the researcher is comfortable with nonsubjectively determined prior odds ratios and using the Hausdorff measure rather than the more commonly employed Lebesgue measure.

Villani (2004) develops an efficient procedure for obtaining the posterior distribution of r using the (proper) Uniform prior on the cointegration space in the linear normalization. He derives closed form solutions for the posterior probabilities of $r = 0$ and $r = p$. The posterior probabilities of $r = 1, ..., p-1$ are computed from the posterior sample of α and B under each cointegration rank utilizing the method developed by Chib (1995).

Asymptotic approximations to the integral with respect to (θ, α) have also been used to obtain $p(r|Data)$. Corander and Villani (2004) use the fractional Bayes approach of O'Hagan (1995) to derive an approximation of the posterior distribution of the cointegration rank jointly with the lag length of the VAR and

[4] Improper priors are those which do not integrate to one. Thus, the integrating constant is not finite and must be selected/justified in some manner. Typically, for estimation of parameters, the integrating constant does not matter and use of improper priors is common. However, when doing model comparison (e.g. using Bayes factors or posterior odds), the integrating constant cannot be ignored and, thus, problems such as Bartlett's paradox can arise with improper priors.

the deterministic trend model. The fractional Bayes approach has an intuitive appeal and has been supported by asymptotic arguments, but has also been criticized by a number of authors (see *e.g.* the discussion of O'Hagan (1995) and Fernández, Ley and Steel (2001)) on the ground that it lacks a proper Bayesian interpretation.

After integrating out (θ, α) to provide an analytical form for $p(\beta, r|Data)$, Strachan and Inder (2004) use a Laplace approximation of the integral with respect to β to obtain and expression for $p(r|Data)$. This approach takes advantage of the Poincaré separation theorem from which we can obtain useful analytical results on the posterior, and utilizes techniques developed in James (1954 and 1969). Chao and Phillips (1999) use the posterior information criteria (PIC) developed in Phillips (1996) and Phillips and Ploberger (1996) to select the modal value for r. Related approaches are the common Schwarz Bayes information criteria (Schwarz (1978)) as a first order approximation to the marginal likelihood, while the PIC and Laplace methods use second order approximations. A limitation of these approaches is that there is less control over the accuracy of the approximation than when using MCMC methods.

5.4 Point estimation of the cointegrating space

In standard inference problems one usually summarizes the posterior distribution with a measure of location, e.g. the posterior mean, median or mode, and a measure of posterior spread, e.g. the posterior standard deviation or interquartile range. The exact choice of location and spread is motivated by decision theory. The optimal posterior summary is the estimate minimizing the posterior expected loss

$$\hat{\theta} = \underset{\tilde{\theta} \in \Theta}{\operatorname{argmin}}\, E[l(\theta, \tilde{\theta})],$$

where $l(\theta, \tilde{\theta})$ is a loss function penalizing the discrepancy between the true parameter value and its estimate. For example, the mean, median and mode are the optimal estimates given a quadratic, absolute value or a 0-1 loss on the discrepancy $\theta - \tilde{\theta}$, respectively (see, e.g., Berger (1985)).

It may be tempting to use the standard point estimates on the unrestricted elements of β, e.g. the posterior mode or median of B in the linear normalization (as noted above, the posterior mean does not exist with some priors) and plugging in these estimates in β to obtain an estimate of the cointegration vectors. Villani (2003) criticizes this practise on two grounds. First, the loss functions underlying the standard estimates are not suitable for point estimation of the cointegration vectors. Since only the cointegration space is identified, a loss function for estimating β should measure the discrepancy between *linear subspaces* rather than between real matrices. It is not true that if two matrices β_1 and β_2 are far apart then $\operatorname{sp}\beta_1$ is distant from $\operatorname{sp}\beta_2$. Optimal estimates of the cointegration vectors should thus be based on distance measures developed for the Grassman manifold. Second, by focusing on estimating unrestricted parameters under a given normalization rule we may end up with estimates

that are not invariant to the chosen normalization. For example, the estimators based on plugging in standard estimates of B into β in the linear normalization are dependent on the order of the variables in the system, *even* if the posterior distribution is invariant.

Strachan and Inder (2003) present a method for estimating the mode of the posterior for the cointegrating space using the Poincaré separation theorem. However, for various reasons many Bayesians are more comfortable with or require mean estimates. Therefore, Villani (2003) derives a more appropriate measure of location using the square of the projective Frobenius linear subspace distance

$$l(\beta, \tilde{\beta}) = \left\| \beta\beta' - \tilde{\beta}\tilde{\beta}' \right\|^2,$$

as loss function, where $\|A\| = tr(A'A)^{1/2}$ is the Frobenius norm for matrices and we have without loss of generality assumed that β is orthonormal. The optimal estimate given the projective Frobenius loss is the matrix of eigenvectors of $E(\beta\beta')$ corresponding to the r largest eigenvalues. Note that $\beta\beta'$ is the projection matrix of the linear subspace spanned by the columns of β. This estimate is thus based on a quadratic loss on the space of projection matrices and may therefore be interpreted as the mean cointegration space, rather than the mean of the unrestricted elements of β.

Villani (2003) proposes a scalar measure of overall uncertainty in the posterior distribution of β based on the projective Frobenius linear subspace distance. He shows that this measure is independent of both the number of variables in the system and the number of cointegrating relations. It may therefore be used to compare the uncertainty regarding the estimated cointegration space in different studies. The measure is bounded between zero and one, where the upper bound is obtained if the posterior distribution is the Uniform distribution on $\mathbb{G}_{p,r}$.

6 Bayesian model averaging in cointegration models

As researchers usually have a range of models to use for their analysis and there is usually uncertainty as to which model generated the data, a growing area of interest in econometrics is the treatment of model uncertainty. Traditionally, choice of the model form has been based on some criteria such as goodness-of-fit or on a series of nested or non-nested hypothesis tests designed to discriminate between alternative models. Inference is then based upon this chosen best model. One problem with this practice is that, once a particular model has been chosen, the fact that a number of other models have been discarded is usually ignored in measures of uncertainty for the object of interest. No allowance is made for the possibility of sample statistics yielding an incorrect choice and assessment of the precision of estimation via standard errors makes no provision for the preliminary-test implications for inference. Accounting for, and incorporating, model uncertainty can be achieved quite naturally

in the Bayesian approach by using Bayesian model averaging (BMA). A useful example of BMA which uses averaging over a range of cointegration models to produce improved forecasts is presented in Villani (2001).

The basic ideas underlying BMA can be demonstrated if we let M_i for $i = 1, \ldots, M$ denote the ith model with parameters λ_i. Suppose we wish to learn about a function of interest common to all models (e.g. an impulse response) $\omega = g(\lambda_i)$. If we treat the M_i as random variables, then the rules of probability imply that:

$$p(\omega|Data) = \sum_{i=1}^{M} p(M_i|Data)\, p(\omega|Data, M_i).$$

In words, inferences about ω should be based on a weighted average of the posteriors in each model, where the weights are given by the posterior model probabilities[5] $p(M_i|Data)$.

Within the framework of cointegrated ECMs a number of models can be defined based on the number of cointegrating vectors, lag lengths, or deterministic terms. We may also have different specifications of cointegrating spaces which we may consider important. For instance, in an application involving foreign and domestic prices and an exchange rate, the theory of purchasing power of parity suggests that a particular cointegrating vector should be present. Such a model implies a restricted version of $\mathfrak{p}$ such as $\mathfrak{p} = sp(H_1)$. Alternatively, one may not wish to impose the theory of purchasing power parity and express ignorance about $\mathfrak{p}$ through the Uniform priors discussed in Subsection 4.2. Other restricted versions of the cointegrated ECM arise if exogeneity is present (Engle, Hendry and Richard (1983) and Urbain (1992)). Although exogeneity has statistical implications, it has also been given economic implications (e.g., Garratt et al. (2003)).

Given a set of models, the implementation of BMA in cointegrated ECMs is straightforward, requiring only a posterior simulator for every model (see the discussion in Section 5) and a method for calculating $p(M_i|Data)$ (see the discussed in subsection 5.3). In a comprehensive study in which a range of methods were used, Strachan and van Dijk (2004) present posterior probabilities for very large classes of models covering all the issues we have mentioned so far: cointegrating rank; lag length; deterministic processes; overidentifying restrictions; and weak exogeneity. The empirical results suggest that BMA is an important issue in that results are substantially different from those obtained from a single model. This is undoubtedly an issue of empirical importance that will provoke a great deal of future research.

[5] The posterior model probability is the prior model probability times the marginal likelihood. For cointegrated ECM's there are many ways to obtain the marginal likelihood. The discussion of subsection 5.3 implicitly describes some of these.

7 Conclusion

In this survey, we have discussed Bayesian inference in the cointegrated error correction model with a focus in particular on r (the number of cointegrating vectors) and the cointegration space. A strong message of this survey is that, in light of global and local non-identification problems, prior elicitation is an important issue and that it is better to think in terms of the cointegration space than in terms of cointegrating vectors. In particular, apparently "noninformative" priors on cointegrating vectors can imply very (unreasonably) informative priors for the cointegration space and can even result in improper posteriors. In light of this, we have discussed in detail two main approaches to prior elicitation which surmount some or all of the problems caused by the identification problems. These approaches, which have been found to yield sensible inferences, we have called the Embedding Approch and the Grassman Approach.

Fortunately, even using the Embedding and Grassman approaches, posterior computation is relatively straightford and simple MCMC algorithms can be derived. Accordingly, this paper has only briefly discussed posterior computation.

As the majority of studies in the literature have use the traditional linear cointegrated error correction model, this has been our focus in this survey. A noteable exception is Martin and Martin (2000) who present a method of obtaining inference on the triangular cointegration model of Phillips (1991) using the Jeffreys' prior. However, the models and methods we have described should also be useful when it comes to considering a myriad of possible extensions to the ECM (e.g. threshold/Markov switching cointegration models, ECMs with structural breaks or time varying parameters, different error structures, etc.). Bayesian prior elicitation and posterior computation methods typically can be interpreted in terms of blocks of parameters. Many extensions can be interpreted as adding a new block of parameters to our current set. All the issues of prior elicitation and posterior computation described in this survey are relevant for the appropriate block of the extended model. As an example, collect all of the parameters $r, \alpha, \beta, \Psi, \Phi$ and Σ into θ and consider our ECM, which is parameterized in terms of θ. We have discussed methods for eliciting the prior $p(\theta)$ and simulating from the posterior $p(\theta|Data)$. Now suppose we wish to extend this model to allow for a single structural break at unknown time, T^*, with θ_j for $j = 1, 2$ denoting the ECM parameters before and after the break. Bayesian prior elicitation would typically involve specifying $p(\theta_j)$ and $p(T^*)$. But the issues and approaches described in this survey are relevant for $p(\theta_j)$ and, thus, the researcher need only worry about $p(T^*)$. Posterior simulation would typically involve drawing from $p(\theta_j|Data, T^*)$ (or $p\left(r_j, \beta_j|Data, T^*\right)$) and $p(T^*|Data)$ (or some posterior conditional such as $p(T^*|Data, \theta_1, \theta_2)$). But the discussion in this survey is relevant for $p(\theta_j|Data, T^*)$ and the researcher need only worry about developing a method for simulating from $p(T^*|Data)$.

Bayesian work on extensions of the cointegrated ECMs is still in its infancy and is a promising area for future research. One extension worth noting is by Paap and van Dijk (2003), who consider a Markov switching extension of a cointegrated ECM. Another is by Martin (2001) who considers a fractionally

cointegrated model. Finally, a recent PhD thesis (Sugita (2004)) considers both a Markov switching and a structural break extension of the cointegrated ECM.

8 Acknowledgements.

The authors would like to thank Luc Bauwens and Frank Kleibergen for useful comments that have improved this chapter.

9 References

Bauwens, L. and A. Giot, 1998, A Gibbs sampling approach to cointegration, Computational Statistics 13, 339-368.

Bauwens, L. and M. Lubrano, 1996, Identification restrictions and posterior densities in cointegrated Gaussian VAR systems, in: Advances in Econometrics 11, Part B, (JAI Press, Greenwich) 3-28.

Bauwens, L., M. Lubrano and J.-F. Richard, 1999, Bayesian inference in dynamic econometric models (Oxford University Press, Oxford).

Bauwens, L. and J.-F. Richard, 1985, A 1-1 poly-t random variable generator with application to Monte Carlo integration, Journal of Econometrics 29, 19–46.

Bauwens L. and H.K. van Dijk, 1990, Bayesian limited information analysis revisited, in: J. Gabszewicz, J.F. Richard and L. Wolsey, eds., Economic decision-making: Games, econometrics, and decision-making, Contributions in honour of Jacques Drèze (North-Holland, Amsterdam) 385-424.

Berger, J. O., 1985, Statistical decision theory and Bayesian analysis, second edition (Springer Verlag, New York)

Boswijk, H.P., 1996, Testing identifiability of cointegrating vectors, Journal of Business and Economic Statistics 14, 153-160.

Chao J. and P.C.B. Phillips, 1999, Model selection in partially nonstationary vector autoregressive processes with reduced rank structure, Journal of Econometrics 91, 227-271.

Chib, S. and E. Greenberg, 1995, Understanding the Metropolis-Hastings algorithm, The American Statistician 49, 327-335.

Corander, J. and Villani, M., 2004, Bayesian assessment of dimensionality in reduced rank regression, forthcoming in Statistica Neerlandica.

DeJong, D., 1992, Co-integration and trend-stationarity in macroeconomic time series, Journal of Econometrics 52, 347-370.

Dorfman, J., 1994, A numerical Bayesian test for cointegration of AR processes, Journal of Econometrics 66, 289-324.

Drèze, J.H., 1978, Bayesian regression analysis using poly-t densities, Journal of Econometrics 6, 329-354.

Engle, R.F. and C.W.J. Granger, 1987, Co-integration and error correction: Representation, estimation and testing, Econometrica 55, 251-276.

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

Escobar, M. and M. West, 1995, Bayesian density estimation using mixtures, Journal of the American Statistical Association 90, 577-588.

Fernández, C., E. Ley and M. F. J. Steel, 2001, Benchmark priors for Bayesian model averaging, Journal of Econometrics 100, 381-427.

Garratt, A, K. Lee, M.H. Pesaran, and Y. Shin, 2003, A Long run structural macroeconomic model of the UK, Economic Journal 113, 412-455.

Geweke, J., 1993, Bayesian treatment of the independent Student-t linear model, Journal of Applied Econometrics 8, S19-S40.

Geweke, J., 1996, Bayesian reduced rank regression in econometrics, Journal of Econometrics 75, 121-146.

James, A.T., 1954, Normal multivariate analysis and the orthogonal group, Annals of Mathematical Statistics 25, 40-75.

James, A.T., 1969, Test of equality of the latent roots of the covariance matrix, in: P.R. Krishnaiah, ed., Multivariate analysis, *Vol.* II (Academic Press, New York) 205–218.

Johansen, S., 1995, Likelihood-based inference in cointegrated vector autoregressive models (Oxford University Press, Oxford).

Khatri, C. G. and K. V. Mardia, 1977, The von Mises-Fisher distribution in orientation statistics, Journal of Royal Statistical Society B 39, 95–106.

Kleibergen F., 1997, Bayesian simultaneous equations analysis using equality restricted random variables, 1997 Proceedings of the Section on Bayesian Statistical Science, 141-147, American Statistical Association.

Kleibergen F., 2004, How to overcome the Jeffreys'-Lindley paradox, forthcoming in Journal of Econometrics.

Kleibergen, F. and R. Paap, 2002, Prior, posteriors and bayes factors for a Bayesian analysis of cointegration, Journal of Econometrics 111, 223-249.

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

Kleibergen, F. and H.K. van Dijk, 1998, Bayesian simultaneous equations analysis using reduced rank structures, Econometric Theory 14, 701-743.

Kleibergen, F. and E. Zivot, 2003, Bayesian and classical approaches to instrumental variable regression, Journal of Econometrics 114, 29-72.

Kloek, T. and H. K. van Dijk, 1978, Bayesian estimats of equation system parameters: an application of integration by Monte Carlo, Econometrica 46, 1-19.

Koop, G., 1991, Cointegration tests in present value relationships: A Bayesian look at the bivariate properties of stock prices and dividends, Journal of Econometrics 49, 105-140.

Koop, G., 1994, An objective Bayesian analysis of common stochastic trends in international stock prices and exchange rates, Journal of Empirical Finance 1, 343-364.

Litterman, R., 1986, Forecasting with Bayesian vector autoregressions: Five years of experience, Journal of Business and Economic Statistics 4, 25-38.

Luukkonen, R., A. Ripatti and P. Saikkonen, 1999, Testing for a valid normalisation of cointegrating vectors in vector autoregressive processes, Journal of Business and Economic Statistics 17, 195-204.

Martin, G., 2001, Bayesian analysis of a fractional cointegration model, Econometric Reviews 20, 217-234.

Martin, G. and V. Martin, 2000, Bayesian inference in the triangular cointegration model using a Jeffreys prior, Communications in Statistics, Theory and Methods 29, 1759-1785.

Nelson, C. and C. Plosser, 1982, Trends and random walks in macroeconomic time series: Some evidence and implications, Journal of Monetary Economics 10, 139-162.

O'Hagan, A., 1995, Fractional Bayes factors for model comparison, Journal of the Royal Statistical Society B 57, 99-138.

Paap, R. and H.K. van Dijk, 2003, Bayes estimates of Markov trends in possibly cointegrated series: An application to US consumption and income, Journal of Business and Economic Statistics 21, 547-563.

Phillips, P.C.B., 1991, Optimal inference in cointegrated systems, Econometrica 59, 283-306.

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

Phillips, P.C.B., 1996, Econometric model determination, Econometrica 64, 763-812.

Phillips, P.C.B. and W. Ploberger, 1996, An asymptotic theory of Bayesian inference for time series, Econometrica 64, 381-412.

Richard, J.-F. and H. Tompa, 1980, On the evaluation of poly-t density functions, Journal of Econometrics 12, 335-351.

Schwarz, G. ,1978, Estimating the dimension of a model, Annals of Statistics 6, 461-464.

Strachan, R. 2003, Valid Bayesian estimation of the cointegrating error correction model, Journal of Business and Economic Statistics 21, 185-195.

Strachan, R. and B. Inder, 2004, Bayesian analysis of the error correction model, forthcoming in Journal of Econometrics.

Strachan, R. and H.K. van Dijk, 2004, Valuing structure, model uncertainty and model averaging in vector autoregressive processes, Econometric Institute Report EI 2004-23, Erasmus University Rotterdam.

Strachan, R and H.K. van Dijk, 2003, Bayesian model selection for a sharp null and a diffuse alternative with econometric applications, Oxford Bulletin of Economics and Statistics 65, 681-688.

Sugita, K., 2004, Bayesian analysis of cointegrated vector autoregressive models, PhD thesis, Warwick University.

Urbain, J. P., 1992, On weak exogeneity in error correction models, Oxford Bulletin of Economics and Statistics 52, 187-202.

Verdinelli, I. and L. Wasserman, 1995, Computing Bayes Factors using a generalization of the Savage-Dickey Density Ratio, Journal of the American Statistical Association 90, 614-618.

Villani, M., 2000, Aspects of Bayesian cointegration, PhD thesis, Stockholm University.

Villani, M., 2001, Bayesian prediction with cointegrated vector autoregressions, International Journal of Forecasting 17, 585-605.

Villani, M., 2003, Bayes estimators of the cointegration space, Riksbank Working Paper No. 150, September 2003.

Villani, M., 2004, Bayesian reference analysis of cointegration, forthcoming in Econometric Theory.

Zellner, A., 1971, An introduction to Bayesian inference in econometrics (John Wiley and Sons Inc., New York).