Improper priors with well defined Bayes Factors.

Econometric Institute Report EI 2004-18

Rodney W. Strachan[1]
Herman K. van Dijk[2]
[1]Department of Economics and Accounting,
University of Liverpool,
Liverpool, L69 7ZA, U.K.
email: rstrach@liverpool.ac.uk
[2]Econometric Institute,
Erasmus University Rotterdam,
Rotterdam, The Netherlands
email: hkvandijk@few.eur.nl

April 21, 2004

ABSTRACT

A sensible Bayesian model selection or comparison strategy implies selecting the model with the highest posterior probability. While some improper priors have attractive properties such as, e.g., low frequentist risk, it is generally claimed that Bartlett's paradox implies that using improper priors for the parameters in alternative models results in Bayes factors that are not well defined, thus preventing model comparison in this case. In this paper we demonstrate this latter result is not generally true and expand the class of priors that may be used for computing posterior odds to include some improper priors. Our approach is to give a new representation of the issue of undefined Bayes factors and, from this representation, develop classes of

improper priors from which well defined Bayes factors may be derived. This approach involves either augmenting or normalising the prior measure for the parameters. One of these classes of priors includes the well known and commonly employed shrinkage prior. Estimation of Bayes factors is demonstrated for a reduced rank model.
Key Words: Improper prior; Bayes factor; marginal likelihood; shrinkage prior; measure.
JEL Codes: C11; C52; C15; C32.

1 Introduction.

Accounting for uncertainty associated with stochastic variables and parameter estimates is a central issue in statistical analysis. A natural extension is to account for uncertainty associated with the statistical or economic model of the process used in the analysis. A typical approach to data analysis is to select the 'best' of a set of competing models and then condition upon a particular model and ignore the uncertainty associated with that model. An attractive feature of the Bayesian approach to inference is the natural way in which model uncertainty may be assessed and incorporated into the analysis via the posterior probabilities of the models. An important method of incorporating this uncertainty that has attracted much attention in recent years is through Bayesian model averaging (BMA). In either BMA or Bayesian model selection the important entity is the posterior probabilities of the models in the model set.

Another attractive feature of Bayesian analysis is the ability to incorporate into the analysis the prior distribution. This brings advantages such as allowing the researcher to reflect in the analysis a range of prior beliefs - from ignorance to dogma - that may reflect personal preferences, or that have justifications on decision-theoretic grounds, or that result in improved estimation performance. However, since Bartlett (1957) it has generally been accepted, even occasionally explicitly stated, that improper priors on all of the parameters of one or more models result in ill-defined Bayes factors and in posterior probabilities that prefer (with probability one) the smaller model regardless of the information in the data. This is commonly termed Barlett's paradox.

As a result of Bartlett's paradox, the principle is generally adopted when computing posterior model probabilities that the use of improper priors should be restricted to the common parameters (such as the variance and perhaps the intercept in a linear model) and proper priors must be specified for the remaining parameters. The adoption of this principle has precluded the use of improper priors in model comparison and model selection where posterior probabilities play an important role. The perceived problem is, essentially, that a well-defined posterior distribution for the model set is not obtainable. This result contrasts with analysis of the parameters in a particular model for which well defined posteriors can be obtained even with improper priors on some or all parameters. This is a well understood result. However, as the model may be considered by a Bayesian as just another parameter, it seems incongruous - even paradoxical - that the use of improper priors does not also result in a well defined posterior distribution for the model. In this paper we show that this is a paradox only and not an inconsistency, and does not hold generally.

Our aim is to widen the class of priors that may be used to obtain posterior probabilities for use in such exercises as BMA, to include some improper priors[1]. To further this end, we demonstrate that Bartlett's paradox does not hold for all improper priors - contrary to conventional wisdom. We provide a new representation of the issue by decomposing the differential term for the vector of parameters in a model into one term that is a measure for a compact space such that it results in a finite term in the normalising constant of the prior, and a second term defined on the unbounded space such as $R^+ = \{x : x \geq 0\}$, which results in an infinite prior measure for this space. We use this representation in two ways. First, this representation allows us to develop a new prior that results in well defined Bayes factors and has properties similar to some priors already in use. Second, we demonstrate that the improper shrinkage prior - which has good properties in terms of frequentist risk (see for example Ni and Sun, 2003) - also results in well defined Bayes factors.

Much of the earlier and some of the more recent literature on BMA has focused upon the Normal linear regression with uncertainty in the choice of regressors (for a good introduction to this large body of literature, see Fernandes, Ley and Steel 2001). Another contribution of this paper is to extend the class of models and problems that may be considered with BMA.

[1]We emphasise that it is not the aim of this paper to produce another method of obtaining inference on model uncertainty that may be regarded as objective or as a reference approach. This research takes a different direction.

For much of the discussion we leave the form of alternative models largely unspecified except for their dimension. We demonstrate application of the priors discussed in a cointegrating vector error correction model (VECM) in which we consider the uncertainty associated with numbers of stochastic trends, type of deterministic processes, identification of the cointegrating space and lag length.

As already mentioned, improper priors play an important role in Bayesian analysis not just because they are convenient and commonly employed representations of ignorance. Some of these priors have information theoretic justifications such as the Jeffreys' prior, while others result in estimators that are better in some sense such as having lower frequentist risk. This latter result is important for exercises such as forecasting or impulse response analysis. Being able to use some of these priors when calculating posterior model probabilities would therefore allow us to retain these benefits.

The structure of the paper is as follows. In Section 2 we outline the explanation for why the posterior distribution is well defined when a flat prior for the parameters with unbounded support is employed, while the Bayes factors are not. This also explains why improper priors on common parameters can be employed in estimating posterior probabilities of the models. This is already a reasonably well understood issue, but we present it using the decomposition of the differential term to motivate the approach in the rest of the paper. In Section 3 we discuss other approaches to obtaining inference with improper priors that have been presented in the literature. The improper priors are developed in Section 4 and, in Section 5, are applied to some well known theoretical examples and to an empirical example relating to the term structure of Australian interest rates. Section 6 contains some concluding comments and suggestions for further research.

First we must introduce some notation for matrix spaces and measures on these spaces for use in developing the discussion. For further discussion of these concepts see Strachan and Inder (2004) and Strachan and van Dijk (2004). The $r \times r$ orthogonal matrix C is an element of the orthogonal group of $r \times r$ orthogonal matrices denoted by $O(r) = \{C(r \times r) : C'C = I_r\}$, that is $C \in O(r)$. The $n \times r$ semi-orthogonal matrix V is an element of the Stiefel manifold denoted by $V_{r,n} = \{V(n \times r) : V'V = I_r\}$, that is $V \in V_{r,n}$. If $r = 1$, then V is a vector which we will denote by lower case such as v and $v \in V_{1,n}$. Finally, let $\lambda(A)$ denote the Lebesgue of the collection of spaces A, and $\lambda(A) = \infty$ to denote that A has infinite Lebesgue measure.

2 The posterior and Bartletts' paradox.

In this section we restrict ourselves to the uniform prior as used in Bartlett's original example as this is sufficient to demonstrate the methods we use and motivate the later derivations.

Consider the investigation of the properties of a vector of data y. Let the i^{th} model in a model set be denoted by M_i, $i = 1, 2, \ldots$ and the n_i vector of parameters for this model as θ_i. The posterior probability of the model given by $P(M_i|y)$ is a useful measure of the support in y for M_i. The ratio of the posterior probabilities for two models is proportional to the Bayes factor for these two models, B_{ij}, and if the two models are considered *a priori* equally likely, this ratio is equal to the Bayes factor. It is through this relationship with the Bayes factor (i.e. by assuming equal prior model probabilities) that the posterior probabilites are most often obtained. Our aim in this section is to provide an alternative representation of Bartlett's paradox. However, we begin with a discussion of the definition of the posterior with improper priors as this explanation is well understood, generally accepted[2], and leads directly to an understanding of why some improper priors result in well defined Bayes factors.

Let the n_i vector of parameters θ_i have support defined by $\theta_i \in \Theta_i \subseteq R^{n_i}$ with $\lambda(\Theta_i) = \infty$. We ignore parameters with compact supports with finite Lebesgue measure as they do not cause problems with the interpretation of the Bayes factor. Therefore when we refer to a model having a particular dimension, we intend by this the dimension of the space Θ_i of the model. If the prior density on θ_i is $\pi_i(\theta_i) = h_i(\theta_i)/\mathfrak{c}_i$ where $\mathfrak{c}_i = \int h_i(\theta_i)\, d\theta_i$, and the likelihood function is $L_i(\theta_i)$, the posterior density is defined as

$$\begin{aligned}
\pi_i(\theta_i|y) &= \frac{L_i(\theta_i)\,\pi_i(\theta_i)}{\int_{\Theta_i} L_i(\theta_i)\,\pi_i(\theta_i)\,d\theta_i} \\
&= \frac{L_i(\theta_i)\,h_i(\theta_i)/\mathfrak{c}_i}{\int_{\Theta_i} L_i(\theta_i)\,h_i(\theta_i)/\mathfrak{c}_i d\theta_i} \\
&= \frac{L_i(\theta_i)\,h_i(\theta_i)}{\int_{\Theta_i} L_i(\theta_i)\,h_i(\theta_i)\,d\theta_i}.
\end{aligned}$$

Notice that even if we use an improper prior such as with $h_i(\theta_i) = 1$ and $\lambda(\Theta_i) = \infty$ such that $\mathfrak{c}_i = \infty$, the posterior is well defined so long as the

[2]Although we ackowledge that there remain some issues with the resulting posterior - see for example Stone and Dawid (1972).

integral $p_i = \int_{\Theta_i} L_i(\theta)_i \, h_i(\theta_i) \, d\theta_i$ converges. We assume this is the case throughout the paper such that we only consider proper posteriors.

For comparison of two models M_i and M_j we can use the posterior odds ratio written as

$$\frac{\Pr(M_i|y)}{\Pr(M_j|y)} = \frac{\Pr(M_i)}{\Pr(M_j)} \frac{m_i}{m_j} = \frac{\Pr(M_i)}{\Pr(M_j)} B_{ij}$$

where $B_{ij} = m_i/m_j$ is the Bayes factor (in favour of model i against model j) and $m_i = p_i/\mathfrak{c}_i$ is the marginal density of y under model i. Therefore, $B_{ij} = p_i/p_j \times \mathfrak{c}_j/\mathfrak{c}_i$. As we only consider proper posteriors the ratio p_i/p_j will be well defined. If a proper prior is used for each model such that $\mathfrak{c}_i < \infty$ and $\mathfrak{c}_j < \infty$ are well defined - and possibly known or able to be estimated - the Bayes factor is well defined as the ratio $\mathfrak{c}_j/\mathfrak{c}_i$ is also defined.

If, however, we use an improper prior of the form $h_j(\theta_j) = 1$ with $\lambda(\Theta_j) = \infty$ for M_j and a proper prior for M_i, then $\mathfrak{c}_j$ will be infinite such that the ratio $\mathfrak{c}_j/\mathfrak{c}_i$ is ∞ and so the Bayes factor is not well defined and nor are the posterior probabilities in the sense that their value will not reflect any information in the data. Further, if we use an improper prior of the form $h_k(\theta_k) = 1$ for both $k = 1, 2$, then the ratio $\mathfrak{c}_j/\mathfrak{c}_i$ is either 0, 1 or ∞ depending upon the relative dimensions of the two models and so, in the first and last cases, the Bayes factor is not always well defined and nor are the posterior probabilities. The exception being when $\mathfrak{c}_j/\mathfrak{c}_i = 1$, which is not entirely helpful as it only holds when the dimensions of the models match.

To explore this issue further, we assume $\Theta_i \equiv R^{n_i}$ and use the decomposition of the $n_i \times 1$ vector θ_i into $\theta_i = v_i\tau$ where the $n_i \times 1$ vector v_i is the polar part and defines the direction of the vector and $\tau > 0$ defines the vector length. The vector for the polar component is of unit length $v_i'v_i = 1$ and is therefore defined as an element of a Stiefel manifold V_{1,n_i}, $v \in V_{1,n_i}$. The compact space V_{1,n_i} has a measure defined by $dv_1^{n_i} = k = 2\overset{n_i}{\Lambda} v'_{i,k}dv$ where $V = [v_i, v_{i,2}, \ldots, v_{i,n_i}]$, $V'V = I_{n_i}$, and V_{1,n_i} has volume

$$\varpi_{n_i} = \int_{V_{1,n_i}} dv_1^{n_i} = 2\pi^{n_i/2}/\Gamma(n_i/2) < \infty \qquad (1)$$

(Muirhead, 1982). We can therefore decompose the differential term for θ_i into $d\theta_i = \tau^{n_i-1}(d\tau)\, dv_1^{n_i}$.

The expression for the differential term leads to the following explanation for Bartlets' paradox and therefore why, although the posterior is well defined, the Bayes factors are not well defined when we use improper priors and models of different dimension. Using the above decomposition of the differential term we can decompose the integral $\mathfrak{c}_i$ into a convergent (finite) part, ϖ_{n_i}, and the divergent part, α_{n_i}. That is,

$$\mathfrak{c}_i = \int_{R^{n_i}} d\theta_i = \int_{R^+} \tau^{n_i - 1} (d\tau) \int_{V_{1,n_i}} dv_1^{n_i} = \alpha_{n_i} \varpi_{n_i} \tag{2}$$

where

$$\alpha_{n_i} = \int_{R^+} \tau^{n_i - 1} (d\tau) = \infty. \tag{3}$$

Note that the integrals α_{n_i} and ϖ_{n_i} do not depend upon the chosen model, only its dimension, n_i.

Next consider an n_j dimensional model with parameter vector $\theta_j = v_j \tau$ with differential term $d\theta_j = \tau^{n_j - 1} (d\tau) \, dv_1^{n_j}$ and, similarly, with integrals

$$\mathfrak{c}_j = \int_{R^{n_j}} d\theta_j = \int_{R^+} \tau^{n_j - 1} (d\tau) \int_{V_{1,n_j}} dv_1^{n_j} = \alpha_{n_j} \varpi_{n_j}.$$

Recall that the posterior is well defined even if the integral $\mathfrak{c}_j = \int_{R^{n_j}} h_j(\theta_j) \, d\theta_j$ does not converge because the integrals in the numerator and denominator diverge at the same rate such that their ratio is one. This same reasoning implies that if $n_i = n_j$ and $h_i(\theta_i) = h_j(\theta_j) = 1$, then the Bayes factor

$$\begin{aligned} B_{ij} &= m_i / m_j = p_i / p_j \times \mathfrak{c}_j / \mathfrak{c}_i \\ &= p_i / p_j \times \alpha_{n_j} / \alpha_{n_i} \times \varpi_{n_j} / \varpi_{n_i} = p_i / p_j \end{aligned}$$

is well defined since $\alpha_{n_i} = \alpha_{n_j}$ and $\varpi_{n_i} = \varpi_{n_j}$ and so $\mathfrak{c}_j / \mathfrak{c}_i = 1$. This result does not require that the models nest, simply that they be of the same dimension, or at least that the number of parameters with supports with infinite Lebesgue measure are the same. When $n_j > n_i$, the integrals of τ (the term α_n) diverge at different rates. That is $\int_0^x \tau^{n_j - 1} (d\tau) > \int_0^x \tau^{n_i - 1} (d\tau)$, such that the ratio $\alpha_{n_j} / \alpha_{n_i} = \infty$. The values of these ratios can be derived by treating them as limits and using l'Hopital's rule and the Radon-Nikodym derivative. The term in B_{ij} due to the polar part will always be finite and known with value

$$\varpi_{n_j} / \varpi_{n_i} = \pi^{(n_j - n_i)/2} \frac{\Gamma(n_i/2)}{\Gamma(n_j/2)}. \tag{4}$$

However, the Bayes factor B_{ij} is again undefined. More extensive discussion of this issue can be found in, for example, Bartlett (1957), Zellner (1971), O'Hagan (1995), Berger and Perrichi (1996) and Lindley (1997).

3 Approaches suggested in the literature to deal with this problem.

As posterior model probabilities can be sensitive to the prior used, much effort has been devoted in the literature to obtaining inference with objective or reference prior. The aim of this work has generally been to obtain a technique that produces posterior model probabilities that contain no subjective prior information. One important reason for using improper priors such as the uniform prior on the support of the parameters is that this is often seen as representing ignorance or uninformative prior beliefs such that inference based on this prior may be regarded as objective or as a reference to which inferences can be compared. However, as we have shown, only inference conditional upon a particular model is obtainable with this uniform prior.

A number of authors have suggested that the undefined ratio $\mathfrak{c}_j/\mathfrak{c}_i$ may be replaced with estimates based upon some minimal amount of information from the sample. Examples of such approaches are Spiegelhalter and Smith (1982), O'Hagan (1995), and Berger and Pericchi (1996). This approach has an intuitive appeal and has been supported by asymptotic arguments. However, as discussed in Fernándes, Ley and Steel (2001), the use of the data to attribute a value to $\mathfrak{c}_j/\mathfrak{c}_i$ involves an invalid conditioning such that the posterior cannot be interpreted as the conditional distribution given the data.

An alternative approach that has been proposed which maintains a valid interpretation of the posterior is to use proper priors. The rationale here is to compare Bayes factors for models with the same amount of prior information. To this end, Fernández *et al.* (2001) propose reference priors for the linear regression model which allow such comparison of results. They use improper priors on the common parameters - the intercept and the variance - and a zero mean normal prior on the remaining coefficients. This approach is supported by the argument of Lindley (1997) that only proper priors should be employed to represent uncertainty. Lindley used model comparison as one motivating example.

However, as we have argued, some improper priors have attractive properties and do result in well defined Bayes factors and posterior probabilities. One approach to using improper priors is given in Kleibergen (2004) who uses the Hausdorff measure and Hausdorff integrals rather than the Lebesgue measure and integrals to develop prior probabilities for models and prior distributions for parameters within models nested within an encompassing model. An advantage of this approach is that it can be used with a very general form for the prior, not simply improper priors. A restriction is that prior model probabilities are designed to diverge at such a rate as to offset the divergent behaviour of the Bayes factor, and so we are restricted in choosing prior model probabilities.

It should be noted that the following result does not require models to nest, nor does it place any restriction upon the specification of the prior probabilities for the models and produces valid Bayesian inference. We show how certain improper priors result in well defined Bayes factors independent of the prior odds ratio. This second point is important as it allows us to employ subjective beliefs such as the statement "I believe M_1 is twice as likely to be true as M_2", or $PROR = \Pr(M_1) / \Pr(M_1) = 2$.

4 Improper priors with well defined Bayes factors: Exceptions to Barlett's paradox.

Augmenting the differential term.

The lack of definition of the Bayes factor for models of different dimensions results from the different rates of divergence in the integrals α_{n_k} $k = i, j$, which in turn results from the different dimensions of the two models. One approach to resolving this issue which suggests itself, is to match the dimensions of the models by augmenting the smaller model with a fictitious vector of parameters of appropriate size and to impose a restriction within the differential to achieve a measure for the smaller model. This augmenting does not require the models to nest, nor do we restrict the augmenting parameter in the same way.

To proceed, let the model M have vector of parameters θ of dimension n while M_0 has parameter vector θ_0 of dimension $n_0 = n - n_1$, $n_1 > 0$, such that the difference in the dimensions is n_1. Let $\theta = \{\theta_0', \theta_1'\}$ where θ_1 is a n_1-dimensional vector. The measure for the prior $h(\theta) = 1$ is given in (2) as $\mathfrak{c} = \alpha_n \varpi_n$. To obtain the measure for the model M_0 we give it the vector of parameters θ and impose the restriction $\theta_1 = 0$. This does not require the models to nest nor that the parameters even have the same interpretation. It can be shown that it is not even necessary that the parameter vectors have the same support, simply that they have support with infinite Lebesgue measure.

The restriction $\theta_1 = 0$ can be imposed by restricting the direction of v in the decomposition $\theta = v\tau$. First, define the $n \times n$ orthogonal matrix

$$\begin{aligned} V &= \begin{bmatrix} v & V_\perp \end{bmatrix} \\ v &= \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} \\ V_\perp &= \begin{bmatrix} V_{00,\perp} & V_{01,\perp} \\ V_{10,\perp} & V_{11,\perp} \end{bmatrix} \end{aligned} \tag{5}$$

such that $V'V = I_n$ $(V \in O(n))$ and v_0 is of dimension $n_0 \times 1$, $V_\perp$ is of dimension $n \times (n-1)$, $V_{00,\perp}$ is of dimension $n_0 \times (n_0 - 1)$, and the dimensions of the remaining matrices are thus defined. The differential $(d\theta) = \tau^{n-1}(d\tau)(dv_1^n)$ derives from the exterior product of the elements of the vector

$$\begin{aligned} (d\theta) &= V'(d\theta) = V'v(d\tau) + V'(dv)\tau \\ &= \begin{bmatrix} v'v \\ V_\perp' v \end{bmatrix} (d\tau) + \begin{bmatrix} v'(dv) \\ V_\perp'(dv) \end{bmatrix} \tau \\ &= \begin{bmatrix} (d\tau) \\ V_\perp'(dv)\tau \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & V_\perp' \tau \end{bmatrix} \begin{bmatrix} (d\tau) \\ (dv) \end{bmatrix} \end{aligned}$$

since $V'(d\theta) = |V|(d\theta)$, $|V| = 1$, and $v'(dv) = -(dv)'v = 0$.

To reduce the dimension of model M from n to n_0, we set $v_1 = 0$, which is equivalent to $\theta_1 = 0$. That is, we restrict the direction of the vector θ such that the subvector θ_0 is zero. Since $v'v = 1$ at all points in $V_{1,n}$ including at $v_1 = 0$, then at this point $v_0'v_0 = 1$ and so $v_0 \in V_{1,n_0}$ and will have the matrix orthogonal complement $V_{00,\perp} \in V_{n_0-1,n_0}$. If $\widetilde{V}_\perp$ is any matrix that spans the orthogonal compliment space of v, then using a similar partitioning as for $V_\perp$ in (5), we have at $v_1 = 0$,

$$\widetilde{V}_\perp' v = \begin{bmatrix} \widetilde{V}_{00,\perp}' v_0 + \widetilde{V}_{01,\perp}' v_1 \\ \widetilde{V}_{10,\perp}' v_0 + \widetilde{V}_{11,\perp}' v_1 \end{bmatrix} = \begin{bmatrix} \widetilde{V}_{00,\perp}' v_0 \\ \widetilde{V}_{10,\perp}' v_0 \end{bmatrix} = 0.$$

This implies that at the point $v_1 = 0$, then $\widetilde{V}_\perp = V_\perp \kappa$ for $\kappa \in O(n-r)$ will be an orthogonal rotation of the matrix $V_\perp$ with $V_{10,\perp} = V'_{01,\perp} = 0$ and $V_{11,\perp} = I_{n_1}$. That is, generally, the space spanned by $\widetilde{V}_\perp$ will lie in the n_1 plane passing through the last n_1 co-ordinate axes and so will have the same differential term as $V_\perp$ since for any $\kappa \in O(n-r)$, $|\kappa| = 1$. To see this, consider the simple case where $n = 3$ and $n_0 = 2$. Thus $v = (v_{11}, v_{21}, v_{31})'$ is a vector in a three dimensional space and each element of the vector relates to a different coordinate of the 3-coordinate system. The column vectors in the matrix $V_\perp$ lie in (and define) the plane spanned by all vectors orthogonal to the vector v. The restriction $v_1 = v_{31} = 0$ implies the third coordinate is always zero and so the vector v is now restricted to the two dimensional plane defined by the first two coordinate axis and the matrix $\widetilde{V}_\perp$ now always lies in the plane passing through the third coordinate axis and defined by the matrix

$$V_\perp = \begin{bmatrix} v_{12} & 0 \\ v_{22} & 0 \\ 0 & 1 \end{bmatrix}.$$

This restriction implies that to obtain the differential term we need only employ the matrix $V_\perp$ and, at the point $v_1 = \theta_1 = 0$, we take exterior products of elements of the vector

$$\begin{aligned}
(d\theta_1) &= V'(d\theta_1) = V'v\,(d\tau) + V'(dv)\,\tau \\
&= \begin{bmatrix} v_0'v_0 + v_1'v_1 \\ V'_{00,\perp}v_0 + V'_{01,\perp}v_1 \\ V'_{10,\perp}v_0 + V'_{11,\perp}v_1 \end{bmatrix}(d\tau) + \begin{bmatrix} v'(dv) \\ V_{00,\perp}(dv_0) + V'_{01,\perp}(dv_1) \\ V_{10,\perp}(dv_0) + V'_{11,\perp}(dv_1) \end{bmatrix}\tau \\
&= \begin{bmatrix} (d\tau) \\ V_{00,\perp}(dv_0)\,\tau \\ (dv_1)\,\tau \end{bmatrix} \text{ at } v_1 = 0 \text{ where } V_\perp = \begin{bmatrix} V_{00,\perp} & 0 \\ 0 & I_{n_1} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 0 & V'_{00,\perp}\tau & 0 \\ 0 & 0 & I_{n_1}\tau \end{bmatrix}\begin{bmatrix} (d\tau) \\ (dv_0) \\ (dv_1) \end{bmatrix}
\end{aligned}$$

and obtain $(d\theta)\,|_{\theta_1=0} = \tau^{n-1}(d\tau)(dv_1^n)\,|_{\nu_1=0} = \tau^{n-1}(d\tau)(dv_1^{n_0})$. By conditioning on $(dv_1^n)\,|_{\nu_1=0} = (dv_1^{n_0})$, we thus obtain the measure

$$\begin{aligned}
\mathfrak{c}_0 &= \int_{R^{n_0}} (d\theta)\,|_{\theta_1=0} = \int_{R^+} \tau^{n-1}(d\tau) \int_{V_{1,n_0}} (dv_1^{n_0}) \\
&= \alpha_n \varpi_{n_0}.
\end{aligned}$$

The ratio of the normalising constants $\mathfrak{c}$ and $\mathfrak{c}_0$ for the priors is then

$$\frac{\mathfrak{c}}{\mathfrak{c}_0} = \frac{\alpha_n \varpi_n}{\alpha_n \varpi_{n_0}} = \pi^{n_1/2} \frac{\Gamma(n_0/2)}{\Gamma(n/2)}$$

and the Bayes factor is well defined as $B = p_0/p \times \mathfrak{c}/\mathfrak{c}_0$ such that the posterior probabilities can be obtained.

In the following we develop the prior implied by this augmenting of the differential for the smaller model. The prior for M is $\pi(\theta) = h(\theta)/\mathfrak{c} = 1/\mathfrak{c}$. Under M_0, as $\theta_0 = v_0 \tau$ implies $(d\theta_0) = \tau^{n_0-1}(d\tau)(dv_1^{n_0})$ and $\theta_0' \theta_0 = \tau^2$, the implied prior for M_0 is then

$$\begin{aligned}
\pi(\theta)|_{\theta_1=0}(d\theta)|_{\theta_1=0} &= h(\theta)|_{\theta_1=0}(d\theta)|_{\theta_1=0}/\mathfrak{c}_0 \\
&= \tau^{n-1}(d\tau)(dv_1^{n_0})/\mathfrak{c}_0 \\
&= \tau^{n_1}\tau^{n_0-1}(d\tau)(dv_1^{n_0})/\mathfrak{c}_0 \\
&= (\theta_0'\theta_0)^{n_1/2}(d\theta_0)/\mathfrak{c}_0.
\end{aligned}$$

As it is the difference in the rates of divergence of the integrals with respect to τ (i.e., α_n) that cause the problems with the Bayes factors, a less formal way of arriving at the same prior is to consider the two differential forms

$$\begin{aligned}
(d\theta) &= \tau^{n-1}(d\tau)(dv_1^n) \\
(d\theta_0) &= \tau^{n_0-1}(d\tau)(dv_1^{n_0}).
\end{aligned}$$

Since $n = n_0 + n_1$ and $\theta_0'\theta_0 = \tau^2$, then clearly if in the prior for M_0 we replace $(d\theta_0)$ by

$$\begin{aligned}
(\theta_0'\theta_0)^{n_1/2}(d\theta_0) &= \tau^{n_1}\tau^{n_0-1}(d\tau)(dv_1^{n_0}) \\
&= \tau^{n-1}(d\tau)(dv_1^{n_0}),
\end{aligned}$$

we have the same result.

Note that for the posterior to be proper requires

$$\int_{R^{n_0}} (\theta_0'\theta_0)^{n_1/2} L_0(\theta_0)\, d\theta_0 = q < \infty$$

where q is finite. The convex form of the prior is similar to the form of the Jeffreys' prior for many models and to the prior of Kleibergen and Paap

(2002). Use of these priors also requires existence of a similar function of the parameters.

Normalising the differential term: Shrinkage priors.

The above augmentation of the differential term results in an improper prior which produces well defined Bayes factors, however it would be reasonable to argue that the implied measure of this approach does not seem a very natural one. We demonstrate in this subsection that an alternative and theoretically more acceptable improper prior is the shrinkage prior advocated and employed by several authors (see for example Stein 1956, 1960, 1962, Lindley 1962, Lindley and Smith 1972, Sclove 1968, 1971, Zellner and Vandaele 1974, Berger 1985, Judge *et al.* 1985, Mittelhammer *et al.* 2000, and Leonard and Hsu 2001).

An important feature of this prior is that it tends to produce an estimator with smaller expected frequentist loss than other standard estimators as may result from flat or proper informative priors (see for example, Zellner 2002 and Ni and Sun 2003). Ni and Sun (2003) provide evidence of this improved performance for estimating the parameters of a VAR and the impulse response functions from these models. Although this prior does not appear to have been considered for model comparison by posterior probabilities, as we now show, it will produce well defined Bayes factors.

The form of the shrinkage prior is

$$\|\theta\|^{-(n-2)} = (\theta'\theta)^{-(n-2)/2}.$$

To demonstrate our claim that the Bayes factor will be well defined, we again use the decomposition $\theta = v\tau$ such that $(\theta'\theta)^{1/2} = \tau$. Thus differential form of the prior is

$$\begin{aligned} (\theta'\theta)^{-(n-2)/2} (d\theta) &= \tau^{-(n-2)}\tau^{n-1} (d\tau) (dv_1^n) \\ &= \tau (d\tau) (dv_1^n) \end{aligned}$$

and this form holds for all models. The normalising constant for model M_i of dimension n is then

$$\begin{aligned} \mathfrak{c}_i &= \int_{R^n} (\theta'\theta)^{-(n-2)/2} (d\theta) = \int_{R^+} \tau (d\tau) \int_{V_{1,n}} (dv_1^n) \\ &= \alpha_2 \varpi_n \end{aligned}$$

such that the ratio of the normalising constants for the shrinkage priors for models of different dimensions is always finite and well defined as the same term α_2 in the normalising constants cancel. Consider two models - the first model M_i with dimension n_i and the second M_j with dimension n_j. The Bayes factor for comparison of the two models with the shrinkage priors will contain the ratio of the normalising constants in the priors. This ratio will be $\varpi_{n_j}/\varpi_{n_i}$ which is given in (4) and is finite and known.

5 Applications

We start with two simple theoretical examples that are well known in the literature, i.e., a comparison between a lognormal and an exponential model and simple zero restrictions in a regression model. Next we analyze a case of a reduced rank regression model which is commonly employed in econometrics for the study of cointegration.

Two simple examples: To provide some simple theoretical examples for applying the above priors, we take the example from Cox (1961) of a comparison of the non-nesting log-normal versus exponential models. For an observation y_t we have the two competing models

$$\begin{aligned} M_f &: \; f\left(y_t|\theta\right) = f_t\left(\theta\right) = y_t^{-1}\left(2\pi\theta_2\right)^{-1/2}\exp\left\{-\frac{\left(\ln y_t-\theta_1\right)^2}{2\theta_2}\right\} \\ -\infty &< \; \theta_1 < \infty, \qquad 0 < \theta_2 < \infty, \qquad y_t > 0 \\ M_g &: \; g\left(y_t|\gamma\right) = g_t\left(\gamma\right) = \gamma^{-1}\exp\left\{-y_t/\gamma\right\} \\ \gamma &> \; 0 \qquad y_t > 0 \end{aligned}$$

That some of the parameters are strictly positive is not a problem as their supports have infinite Lebesgue measures. The differentials for the vectors of parameters are $(d\theta) = \tau\left(d\tau\right)\left(dv_1^2\right)$ and $(d\gamma) = (d\tau)$. For this example we consider using the augmentation of the differential to give the following priors:

$$\pi_f\left(\theta\right) = 1/\mathfrak{c}_f \text{ where } \mathfrak{c}_f = \frac{1}{2}\int_{R^+}\tau\left(d\tau\right)\int_{V_{1,2}}\left(dv_1^2\right) = \frac{1}{2}\alpha_2\varpi_2; \text{ and,}$$

$$\pi_g\left(\gamma\right) = \gamma/\mathfrak{c}_g \text{ where } \mathfrak{c}_g = \int_{R^+}\tau\left(d\tau\right) = \alpha_2.$$

Thus the ratio of the normalising constants is $\frac{\mathfrak{c}_g}{\mathfrak{c}_f} = \frac{1}{\varpi_2} = (2\pi)^{-1}$ and the Bayes factor will be well defined in that it will reflect the evidence in the data.

We will use slightly more complicated models as examples using the shrinkage prior. Here we consider the models

$$\begin{array}{llcl} M_1 & y_i & = & \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \varepsilon_i \\ & \alpha & = & (\alpha_0,\ \alpha_1,\ \alpha_2,\ \alpha_3) \in R^4 \text{ and} \\ M_2 & y_i & = & \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{4,i} + \varepsilon_i \\ & \gamma & = & (\gamma_0,\ \gamma_1,\ \gamma_2) \in R^3 \end{array}$$

The shrinkage priors in which the differential terms are normalised are

$$\pi_1(\alpha) = (\alpha'\alpha)^{-1} / \mathfrak{c}_1 \text{ where } \mathfrak{c}_1 = \int_{R^+} \tau\,(d\tau) \int_{V_{1,4}} \left(dv_1^4\right) = \alpha_2 \varpi_4 \text{ and}$$

$$\pi_2(\gamma) = (\gamma'\gamma)^{-1/2} / \mathfrak{c}_2 \text{ where } \mathfrak{c}_2 = \int_{R^+} \tau\,(d\tau) \int_{V_{1,3}} \left(dv_1^3\right) = \alpha_2 \varpi_3.$$

Thus the ratio of the normalising constants entering into the Bayes factor is $\frac{\mathfrak{c}_2}{\mathfrak{c}_1} = \frac{\varpi_3}{\varpi_4} = \pi^{-1/2} \frac{\Gamma(4/2)}{\Gamma(3/2)} = 0.63662.$

If we instead augment the differentials we obtain the priors

$$\pi_1(\alpha) = 1/\mathfrak{c}_1 \text{ where } \mathfrak{c}_1 = \int_{R^+} \tau^3\,(d\tau) \int_{V_{1,4}} \left(dv_1^4\right) = \alpha_4 \varpi_4 \text{ and}$$

$$\pi_2(\gamma) = (\gamma'\gamma)^{1/2} / \mathfrak{c}_2 \text{ where } \mathfrak{c}_2 = \int_{R^+} \tau^3\,(d\tau) \int_{V_{1,3}} \left(dv_1^3\right) = \alpha_4 \varpi_3.$$

Thus the ratio of the normalising constants is again $\frac{\mathfrak{c}_2}{\mathfrak{c}_1} = \frac{\varpi_3}{\varpi_4}$.

Empirical example: In this section we investigate the reduced rank regression model which in the study of economic time series has several interesting features which allow us to provide a reasonably complete model set. We set up an empirical application of the priors for a p-dimensional time series vector, $y_t = \{y_{jt}\}$ for $j = 1, ..., p$. The results are reported at the end of this section. In the application we use, $p = 4$ and the data for this example is $T = 94$ monthly observations of the 5 year and 3 year Australian Treasury Bond (Capital Market) rates and the 180 day and 90 day Bank Accepted Bill (Money Market) rates from July 1992 to April 2000. This data was previously analyzed in Strachan (2003), Strachan and van Dijk (2003), and Strachan and van Dijk (2004). Our aim is to investigate a simple model implied by the rational expectations theory for the term structure of interest rates (Campbell and Shiller, 1987) in which interest rates are $I(1)$ while the spreads between rates of different maturity are $I(0)$, thus forming cointegrating relations and implying these rates share one common stochastic trend. Although for these variables we might accept that the cointegrating relations (which could possibly be the spreads) may have non-zero means, we would not expect there to be trends in either the levels or the cointegrating relations.

With a maximum of 5 lags and differencing, we have an effective sample size of 88 observations. The VECM of the $1 \times p$ vector time series process $y_t = (y_{1t}, \ldots y_{pt})$, $t = 1, \ldots, T$, conditioning on the l observations $t = -l + 1, \ldots, 0$, is

$$\begin{aligned} \Delta y_t &= y_{t-1}\beta^{+}\alpha + d_t\mu + \Delta y_{t-1}\Gamma_1 + \ldots + \Delta y_{t-l}\Gamma_l + \varepsilon_t \\ &= z_{0,t} = z_{1,t}\beta\alpha + z_{2,t}\Phi + \varepsilon_t \end{aligned}$$

where $z_{0,t} = \Delta y_t = y_t - y_{t-1}$, and we define $z_{1,t}$, $z_{2,t}$, and Φ in the following discussion. The matrices β^{+} and α' are $p \times r$ and assumed to have rank r. We define $d_t\mu$ shortly.

Common features of economic and statistical interest relating to this model are: the number of lags (l) required to describe the short-run dynamics of the system; the form of the deterministic processes in the system (indexed by d); the number of stochastic trends in the system ($p-r$); and the form of the long-run equilibrium relations or the space spanned by the cointegrating vectors (indexed by o). Parameterisation of models with different l is obvious and in the following paragraphs we explain the parameterisation of models with different d, r and o.

For consideration of the range of deterministic processes, the vector μ is decomposed into $\mu = \mu_1\alpha + \mu_0$ where $\mu_1 = \mu\alpha' (\alpha\alpha')^{-1}$ and $\mu_0 = \mu\alpha'_{\perp} (\alpha_{\perp}\alpha'_{\perp})^{-1} \alpha_{\perp}$ such that μ_1 represents the deterministic processes associated with $y_t\beta^{+}$ and μ_0 represents those for y_t (see Johansen, 1995 Section 5.7 for further discussion). Assuming $d_t = (1, t)$, then for each $j = 0, 1$, $d_t\mu_j = \mu_{j,i} + t\mu_{j,\delta}$. Although a wider range of models are clearly available, the five most commonly considered may be stated as follows, where d denotes the model of deterministic terms at given rank r :

$$\begin{aligned}
(d=1) &\ :\ d_t\mu = \mu_{1,i}\alpha + \mu_{0,i} + (\mu_{1,\delta}\alpha + \mu_{0,\delta})\, t \\
(d=2) &\ :\ d_t\mu = \mu_{1,i}\alpha + \mu_{0,i} + \mu_{1,\delta}\alpha t \\
(d=3) &\ :\ d_t\mu = \mu_{1,i}\alpha + \mu_{0,i} \\
(d=4) &\ :\ d_t\mu = \mu_{1,i}\alpha \\
(d=5) &\ :\ d_t\mu = 0
\end{aligned}$$

These five models imply variously that Δy_t and $y_t\beta^+$ may have a nonzero means or trends or some restriction upon these means and trends. For the interest rate data, we would most likely expect $d = 4$ or $d = 5$, as these models imply no trends but a non-zero mean for the cointegrating relation $(d = 4)$ or no deterministic processes (including no non-zero means) in the system $(d = 5)$. From the above decomposition we may define $z_{1,t} = (d_t, y_{t-1})$, $z_{2,t} = (d_t, \Delta y_{t-1}, \ldots, \Delta y_{t-l})$, $\beta = (\mu_1', \beta^{+\prime})'$ and $\Phi = (\mu_0', \Gamma_1', \ldots, \Gamma_l')'$. Notice that the dimension of μ_1 determines the dimension of β as $(p+2) \times r$ for $d = 1$ or 2, $(p+1)\times r$ for $d = 3$ or 4, or $p \times r$ for $d = 5$. Although we consider all five models in the application, in the following discussion we will assume the last case $(d = 5)$ for simplicity.

The aim of cointegration analysis is essentially to determine the dimension and the direction of the cointegrating space. The dimension is determined by the number of common stochastic trends in the system. Of interest when considering the number of stochastic trends is the coefficient matrix β which is of dimension $p \times r$ and we have $rank\,(\beta\alpha) = r \leq p$. When $0 < r < p$, y_t is cointegrated and contains $(p - r)$ stochastic trends, β is the matrix of cointegration coefficients and α is the matrix of factor loading coefficients or adjustment coefficients.

It is common to specify some form for the long-run relations implied by economic or financial theory involving cointegration by restricting the space spanned by β. This implies that, in addition to estimating the dimension of the cointegrating space, we also wish to explore alternative models for the space spanned by the cointegrating vectors. We therefore compare three models for the spaces of interest. When no restriction is placed upon the space and $\rho = sp\,(\beta)$ is free to vary over all of the Grassman manifold, $G_{r,p-r}$, we denote the model by $o = 1$. For the second set of models $(o = 2)$, we refer to the expectations theory which implies the spreads should enter the cointegrating relations and so we are interested in the model with cointegrating space spanned by $H_2 = (h_{2,1}\ h_{2,2}\ h_{2,3}\ )$ where $h_{2,1} = (1, -1, 0, 0)'$,

$h_{2,2} = (0, 1, -1, 0)'$, and $h_{2,3} = (0, 0, 1, -1)'$. In this model we have $\beta = H_2\varphi$ where φ is $3 \times r$ for $r \in [1, 2, 3]$. As the interest rates come from different markets, market segmentation suggests our third set of models of the cointegrating space ($o = 3$) in which we have spaces of interest spanned by $\beta = H_3\varphi$ where φ is $2 \times r$ for $r \in [1, 2]$ and $H_3 = (h_{2,1}\ h_{2,3}\ )$. The models $o = 2$ and $o = 3$ restrict the cointegrating space to subspaces of the space in $o = 1$.

Finally, we introduce the following terms to simplify the expressions in the posteriors. Let $\widetilde{z}_t = (z_{1,t}\beta \quad z_{2,t})$, and the $(pl + r + 1) \times p$ matrix $B = [\alpha'\ \Phi']'$. The model may now be written as $z_{0,t} = \widetilde{z}_t B + \varepsilon_t$ and the dimensions of B will depend upon l, d and r. The model with l lags, $p - r$ stochastic trends, deterministic process d and restriction on the cointegrating space o will be denoted by $M_{(l,r,d,o)}$. If we wish to discuss the general comparison of models we will use M_i to denote one model and M_j to denote a different model.

The prior for β is uniform on $V_{r,p}$ but we adjust the volume to imply a uniform prior on the cointegrating space $\rho = sp(\beta) \in G_{r,p-r}$ where $G_{r,p-r}$ is the Grassman manifold. The prior then becomes $p(\beta) = \frac{1}{c_r}$ (Strachan and Inder, 2004) where

$$c_r = \pi^{-(p-r)r} \Pi_{j=1}^{r} \frac{\Gamma[(p + 1 - j)/2]}{\Gamma[(r + 1 - j)/2]}.$$

The same prior is employed in all models for the covariance matrix. This is the invariant Jeffreys prior for Σ, $p(\Sigma) \propto |\Sigma|^{-(p+1)/2}$ with normalising constant $c_\Sigma = \int_{\Sigma > 0} |\Sigma|^{-(p+1)/2} (d\Sigma) = \infty$.

The prior for model M_i for the n_i-dimensional vector $b = vec(B)$ ($n_i = (p(l-1) + r + 1)p$) is $p(b) \propto (b'b)^{K_i/2}$ where $K_i = \max(n_h) - n_i$ for the prior using augmentation of the differential and $K_i = -(n_i - 2)$ for the shrinkage prior. The normalising constant for the prior will be $c_{i,b} = \int_{R^n} (b'b)^{K_i} (db)$ and so $c_{i,b}$ equals either $\alpha_q \varpi_n$ where $q = \max(n_h)$ or 2.

To sum up, we have the following models in our model set. The rank parameter is an element of $r \in [0, 1, 2, 3, 4]$, the indicator for the deterministic process $d \in [1, 2, 3, 4, 5]$, the lag length $l \in [0, 1, 2, 3, 4]$, and the indicator for overidentification of cointegrating vectors $o \in [1, 2, 3]$. This gives a total of 375 models. Taking account of observationally equivalent or *apriori* impossible models, we need only compute the marginal likelihoods for some 255 models.

In the remainder of this section, we describe a sampling scheme to enable estimation of the marginal likelihoods up to proportionality. Collect the parameters β, Σ, and B into the vector θ. The posterior has the form proportional to $(b'b)^{K/2} L(\theta) |\Sigma|^{-(p+1)/2} \frac{1}{c_i}$ where $c_i = c_r c_\Sigma c_{i,b} = c_r c_\Sigma \alpha_q \varpi_n$. Thus the marginal likelihood for this model is

$$m_i = \int (b'b)^{K/2} L(\theta) |\Sigma|^{-(p+1)/2} (d\theta) \frac{1}{c_i}.$$

We know the c_r and ϖ_n exactly and the term c_Σ is common to all models and so will cancel in the Bayes factor. Therefore to estimate the Bays factor, we need only obtain an estimate of m_i up to an unknown but common proportional constant. To obtain this, we use the proper density $\pi(\theta) = c_{i,\pi}^{-1} L(\theta) |\Sigma|^{-(p+1)/2} = c_{i,\pi}^{-1} h_\pi(\theta)$ where $c_{i,\pi}$ is the estimable normalising constant $c_{i,\pi} = \int h_\pi(\theta)(d\theta)$. Next we define the expectation of the function $(b'b)^{K/2}$ with repect to this distribution as

$$\mu_i = E_\pi\left[(b'b)^{K/2}\right] = \int (b'b)^{K/2} h_\pi(\theta)(d\theta) \frac{1}{c_{i,\pi}}.$$

Thus we may regard the marginal likelihood, m_i, as proportional to the expectation of $(b'b)^{K/2}$ with respect to the density $\pi(\theta)$ such that $m_i = \mu_i c_{i,\pi}/c_i \propto \mu_i c_{i,\pi} (c_{r_i} \varpi_{n_i})^{-1}$ and the Bayes factor for model i to model j is

$$\begin{aligned} B_{ij} &= \frac{m_i}{m_j} = \frac{\mu_i c_{i,\pi}}{c_i} \frac{c_j}{\mu_j c_{j,\pi}} \\ &= \frac{\mu_i c_{i,\pi}}{c_{r_i} c_\Sigma \alpha_q \varpi_{n_i}} \frac{c_{r_j} c_\Sigma \alpha_q \varpi_{n_j}}{\mu_j c_{j,\pi}} \text{ since } c_i = c_r c_\Sigma \alpha_q \varpi_n \\ &= \frac{\mu_i c_{i,\pi}}{c_{r_i} \varpi_{n_i}} \frac{c_{r_j} \varpi_{n_j}}{\mu_j c_{j,\pi}}. \end{aligned}$$

Next we decompose $c_{i,\pi}$ using the following series of integrals in which the functions h_γ are kernals of a density for the parameters in γ :

$$\begin{aligned} c_{i,\pi} &= \int L(\theta) |\Sigma|^{-(p+1)/2} (d\theta) \\ &= \int\int\int h(b,\Sigma|\beta) h(\beta)(d\beta)(db)(d\Sigma) \\ &= c_{\Sigma,y} \int\int h(b|\beta) h(\beta)(d\beta)(db) \\ &= c_{\Sigma,y} c_{b,y} \int h(\beta)(d\beta) \\ &= c_{\Sigma,y} c_{b,y} c_{\beta,y}. \end{aligned}$$

The first two terms $c_{\Sigma,y}$ and $c_{b,y}$ are known exactly (see, for example, Zellner 1971). An estimate of $c_{\beta,y}$ may be obtained by Laplace approximation of Strachan and Inder (2003) or the MCMC approach of Strachan and van Dijk (2003).

Finally, we estimate μ_i by $\widehat{\mu}_i = \frac{1}{N}\Sigma_{g=1}^{N}\left(b^{(g)\prime}b^{(g)}\right)^{K/2}$ where the $b^{(g)}$ are draws from $\pi(\theta)$. Conditional upon β, the parameters Σ and B have the well known inverted Wishart and Multivariate normal distributions respectively. Draws from these distributions are easily obtained. However, β has a non-standard distribution and so draws must be obtained via candidate densities. Details on one method for obtaining such draws are available in Strachan and van Dijk (2003).

We finish this section with a discussion of the results. The Johansen sequential trace test results in selecting a rank of 2 at the 5% level which agrees with the rank selected by AIC, although BIC prefers a rank of zero. Conditional upon $l = 1$, the classical test accepts both restrictions on the cointegrating space, giving support to $o = 3$. Conditional upon $r = 2$, the classical test accept both restrictions on the cointegrating space, giving support to $o = 3$. Recall that the expectations theory implies the interest rates will share one common stochastic trend and the spreads will be $I(0)$. Therefore, these results suggest there is one too many stochastic trends in this system and that the spreads are not cointegrating relations. The extra stochastic trend may result as the interest rates come from different markets and the expectations relations may not hold when comparing rates from these different markets. However, acceptance of $o = 3$ at $r = 2$ does not support this conclusion, and the evidence for or against the theory is not clear. Using information criteria, we find that conditional upon $l = 1$, the AIC prefers $(d = 4, r = 2)$ while BIC prefers $(d = 5, r = 1)$.

For the Bayesian investigation, we estimate the marginal likelihoods from 10,000 draws obtained using a Metropolis-Hastings Markov Chain Monte Carlo technique. Using the shrinkage prior, the model with estimated posterior probability one is $M_{(d,l,r,o)} = M_{(5,1,3,2)}$. This implies the Bayesian results favour the model with no deterministic terms, no lags of differences, a cointegrating rank of three and in which the matrix H_2 lies in the cointegrating space. This result gives clear support - for this data set - to the main features of the Efficient Market Hypothesis that the interest rates share a single common stochastic trend and the spreads are stationary, with a reasonable description of the deterministic and short-run dynamic structure.

Using the approach in which we augment the measure, we find the model with the highest posterior probability is $M_{(4,1,1,1)}$ with posterior probability $P\left(M_{(4,1,1,1)}|y\right) = 0.812$. This implies there is at most a nonzero mean in the cointegrating relation, one lag and three common stochastic trends. The only other two models to receive support with this prior have posterior probabilities of $P\left(M_{(3,1,2,3)}|y\right) = 0.156$ and $P\left(M_{(5,1,3,2)}|y\right) = 0.032$. The log Bayes factor for $M_{(3,1,2,3)}$ to $M_{(4,1,1,1)}$ is -1.650, while the log Bayes factor for $M_{(5,1,3,2)}$ to $M_{(4,1,1,1)}$ is -3.231. These results imply some support for the model selected using the shrinkage prior - $M_{(5,1,3,2)}$, however, they seem to agree more closely with the classical, information criterian results.

6 Conclusion.

Bayesians have generally felt constrained to using proper priors when obtaining posterior probabilities for models for such purposes as Bayesian Model Averaging (BMA). This is unfortunate, as some improper priors have attractive features which the Bayesian may like to employ in such an BMA exercise. Using a relatively simple and well-understood decomposition of the differential term for a vector of parameters, we have demonstrated that the class of priors for which well defined Bayes factors obtain includes some improper priors. One important class is the shrinkage prior which has been shown to to produce estimates with lower frequentist risk than other approaches and therefore are more likely to be admissible under quadratic loss. It is possible that the class of improper priors that permit valid Bayes factors extends beyond those demonstrated in this paper and to others with other attractive properties. This is a potential area for further investigation.

7 References.

Bartlett, M. S. (1957) 'A comment on D.V.Lindley's statistical paradox', *Biometika*, Vol . 44, pp. 533-534.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis* (2nd ed.).New York: Springer-Verlag.

Berger, J. O. and L. R. Pericchi (1996) 'The intrinsic Bayes factor for model selection and prediction' *Journal of the American Statistical Association,* Vol. 19, pp. 109-122.

Campbell J. Y. and R. J. Shiller (1987) 'Cointegration and tests of present value models' *The Journal of Political Economy,* Vol. 95:5, pp. 1062-1088.

Cox, D. (1961) 'Tests of separate families of hypothesis' *Proceedings of the Forth Berkeley Symposium on Mathematical Statistics and Probability.*

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

Garratt, A, K. Lee, M. H. Pesaran, and Y. Shin (2003) 'A Long Run Structural Macroeconomic Model of the UK', *The Economic Journal*, Vol. 113, 412-455.

Johansen, S. (1995) *Likelihood-based Inference in Cointegrated Vector Autoregressive Models.* New York: Oxford University Press.

Judge G. G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. and Lee, T. (1985) *The Theory and Practice of Econometrics* (2nd ed.). New York: Wiley.

Kleibergen, F. (2004) 'Invariant Bayesian inference in regression models that is robust against the Jeffreys-Lindley's paradox', forthcoming in *Journal of Econometrics.*

Kleibergen, F. and R. Paap (2002) 'Priors, Posteriors and Bayes Factors for a Bayesian Analysis of Cointegration', *Journal of Econometrics* **111**, 223-249.

Leonard, T. and Hsu, J. S. J. (2001) *Bayesian Methods.* Cambridge: Cambridge University Press.

Lindley, D.V. (1962) 'Discussion on Professor Stein's paper', *Journal of the Royal Statistical Society Series B*, 24, 285-287.

Lindley, D.V. and Smith, A.F.M. (1972) 'Bayes estimates for the linear model', *Journal of the Royal Statistical Society Series B*, 34, 1-41.

Lindley D. V. (1997) 'Discussion forum: Some comments on Bayes factors', *Journal of Statistical Planning and Inference*, 61, 181-189.

Mittelhammer, R.C., Judge, G.G., and Miller, D.J. (2000) *Econometric Foundations.* Cambridge: Cambridge University Press.

Muirhead, R.J. (1982) *Aspects of Multivariate Statistical Theory.* New York: Wiley.

Ni, S. X. and D. Sun (2003) 'Noninformative Priors and Frequentist Risks of Bayesian Estimators of Vector-Autoregressive Models' *Journal of Econometrics*, Vol. 115, 159-197.

O'Hagan, A. (1995) 'Fractional Bayes Factors for Model Comparison,' *Journal of the Royal Statistical Society, Series B* **57**, 99-138.

Sclove, S. L. (1968) 'Improved Estimators for Coefficients in Linear Regression', *Journal of the American Statistical Association*, 63, 596-606.

Sclove, S.L. (1971) 'Improved Estimation of Parameters in Multivariate Regression', *Sankhya, Series A*, 33, 61-66.

Spiegelhalter, D. J. and A. F. M. Smith (1982) 'Bayes factors for linear and log-linear models with vague prior information', *Journal of the Royal Statistical Society, Series B* **44,** 377–387.

Stone M. and Dawid A. P. (1972) 'Un-Bayesian implications of improper Bayes inference in routine statistical problems' *Biometrika* 59 369-375.

Strachan, R. W. (2003) 'Valid Bayesian estimation of the cointegrating error correction model', *Journal of Business and Economic Statistics,* Vol . 21, pp. 185-195.

Strachan, R. W. and van Dijk (2003) 'Bayesian Model Selection with an Uninformative Prior', *Oxford Bulletin of Economics and Statistics,* Vol . 65, pp. 863-876.

Strachan, R. W. and B. Inder (2004) 'Bayesian Analysis of The Error Correction Model', forthcoming in *Journal of Econometrics.*

Strachan, R. W., and van Dijk, H. K. (2003) 'The value of structural information in the VAR', Econometric Institute Report EI 2003-17, Erasmus University Rotterdam.

Stein, C. (1956) 'Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution' in Proceedings of the *Third Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1 Berkeley, CA: University of California Press, 197-206.

Stein, C. (1960) 'Multiple Regression', in I. Olkin (ed.), *Contributions to Probability and Statistics in Honor of Harold Hotelling.* Stanford: Stanford University Press.

Stein, C. (1962) 'Confidence Sets for the Mean of a Multivariate Normal Distribution', *Journal of the Royal Statistical Society, Series B* 24, 265-296.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

Zellner, A. (2002) 'Bayesian shrinkage estimates and forecasts of individual and total or aggregate outcomes' mimeo University of Chicago.

Zellner, A. and Vandaele, W.A. (1974) 'Bayes-Stein Estimators for k-means, Regression and Simultaneous Equation Models', in Fienberg, S.E. and Zellner, A., (eds.), *Studies in 21 Bayesian Econometrics and Statistics in Honor of Leonard J. Savage.* Amsterdam: North-Holland, 627-653.