# Bayesian Model Selection for a Sharp Null and a Diffuse Alternative with Econometric Applications

Econometric Institute Report EI 2003-12

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

March 14, 2003

## ABSTRACT

In this paper a potential solution is given to the conflict in Bayesian inference between the desire to employ diffuse priors to represent ignorance and the desire to report proper posterior probabilities for alternative models. Using the concept of Stiefel manifolds, diffuse priors are specified on dimension and direction of subspaces of parameter spaces within the context of a linear regression model and a cointegration model. The approach is illustrated using a CAPM and a term structure of interest rates model.

# 1 Introduction.

In a Bayesian approach to model selection, equal prior probabilities are often assigned to alternative models in order to represent 'objective', 'reference' or 'ignorance' prior beliefs. A flat prior density on the parameters within each model is used for the same reason. Such a flat prior specified on the entire real line is an improper density. While commonly employed improper priors result in a well defined and proper posterior density for the parameters of the well known linear regression model, the resulting posterior probabilities used to compare models are not well defined. The intuition of this result may be explained as follows. Consider the case where one compares the sharp null hypothesis of a mean return of zero on some financial asset with the diffuse alternative hypothesis of a mean return that can take on any positive or negative value. Let the posterior probability of the sharp null of a mean return of zero have some finite positive value. Let the alternative hypothesis start from a range of, say, -3 to +3 percent return, which has some positive probability. When one increases this range to (-30, +30) the posterior probability of the alternative hypothesis is 10 times as low. In the limit when the range is ($-\infty$, $+\infty$) the posterior probability of the alternative will be zero and one will always choose the null hypothesis. So, comparing the posterior probabilities of a sharp null with a diffuse alternative will always lead to choosing the null, regardless of prior probabilities and regardless of information in the data, see Section 2 for details.

This is a manifestation of the Lindleys-Jeffreys paradox, which has been studied extensively in the literature and a range of solutions have been proposed. However, as argued in Berger and Perrichi (1996), these proposed solutions are often pseudo-Bayesian in approach, application specific, propose a particular specification of the prior to 'fix' the issue, or use an informative prior. We propose a strictly Bayesian approach that uses a proper, uninformative prior to provide well defined posterior probabilities of alternative models. Our approach is based on a simple procedure to construct a proper diffuse prior on alternative hypotheses using concepts from vector and matrix spaces, in particular Stiefel manifolds. We show how our approach is related to the orthogonal regression concept in classical analysis.

In this paper we focus on a linear combination of variables. We start with a simple mean(s) problem. Next, we discuss testing for a sharp null in a Capital Assets Pricing Model (CAPM). Our main application is the cointegration model for the term structure of interest rates. We note that our approach is also useful for factor models and other multivariate or matricvariate models where linear combinations of variables are used.

The outline of the paper is as follows. In the next section we outline the problem with using improper priors when the aim is model comparison within a linear regression model. In subsection 2.1 we set up the issue of analysis of cointegration spaces. We do this for two reasons: first, to give a context for application of our proposed approach; second, the analysis of cointegrating spaces presents unique issues for which our approach is particularly well suited. In Section 3 we outline the concept of orthogonal regression and introduce our prior for this model setup. Simple econometric applications are presented in Section 4, including the derivation of well defined Bayes factors for cointegration analysis. Section 5 contains some remarks and suggestions for further research.

## 2 Bayes factors and improper priors in a linear model

One may think of a model as being implied by a particular hypothesis in which some of the parameters in an encompassing model are restricted to specific values, thus nesting the restricted model within the encompassing model. The simplest example is the case where a set of i.i.d. observations on some variable, say returns on a financial asset, is available and one wants to test whether the mean return is zero. A comparison of the encompassing and nested models proceeds from the posteriors for the parameters in these models. Although this description implies that one of the models nests within the other, one of the attractive features of the Bayesian is that this is not a requirement. This nested case, however is attractive for exposition because of its simplicity and is sufficient to incorporate the important issues we wish to discuss. Further, this situation is frequently encountered in practice.

Consider the constant mean process $x_t$, $t = 1, \ldots, T$,

$$x_t - \mu = \varepsilon_t. \tag{1}$$

Consider, next, the relative support for a model with $\mu = 0$ and the model in which this equality does not hold. We denote the model in which $\mu = 0$ as $M_0$ and the model in which $\mu \neq 0$ as $M_1$. We denote by $y$ all of the data on explanatory and response variables in the model. A useful summary of the support for such restrictions is the posterior probability for the model, $\Pr(M_i|y)$, $i = 0, 1$, which is implied by the posterior distributions for the parameters for the various models and the prior probabilities for the models.

To obtain a posterior density for a particular model, which we denote $M_i$, we make use of the data density $f_i(y|\theta_i)$ where this density $f_i$ is known as is the dimension of the parameter vector, $\theta_i$. Let the support of $\theta_i$ under $M_i$ be $\Theta_i$. On this support the parameter vector $\theta_i \in \Theta$ is given a prior density denoted $\pi_i(\theta_i) = c_i h_i(\theta_i)$ where $h_i(\theta)$ is a kernal of a density. The normalising constant $c_i$ is defined by the integral $c_i^{-1} = \int_\Theta h_i(\theta_i)(d\theta_i)$. Thus the posterior is proportional to $f_i(y|\theta_i) h_i(\theta_i)$ and is well defined provided the integral $p_i = \int_\Theta f_i(y|\theta_i) h_i(\theta_i)$ converges.

For our model $M_1$ in (1), we might assume $x_t \sim i.i.d.N(\mu, \sigma^2)$ and $\theta_1 = (\sigma^2, \mu)'$ with support $\mathbb{R}^1 \times \mathbb{R}^+$. With a prior $h_1(\theta_1) = h_{1,\mu}(\mu) h_{1,\sigma^2}(\sigma^2)$, the resultant posterior for $\theta_1$ will be proportional to

$$f_1(y|\theta_1) h_1(\theta_1) = \sigma^{-T} \exp\left\{-\frac{1}{2\sigma^2}\left[\nu s^2 + T(\mu - \overline{x})^2\right]\right\} h_1(\theta_1) p_1^{-1}$$

where $\nu = T - 1$,and $s^2$ and $\overline{x}$ are the least squares estimates of $\sigma^2$ and $\mu$ respectively. Assuming a prior density for $\sigma^2$ of $h_{1,\sigma^2}(\sigma^2) = \sigma^{-1}$, then integrating with respect to $\sigma^2$ we obtain the posterior for $\mu$ in $M_1$ as proportional to

$$\left[\nu s^2 + T(\mu - \overline{x})^2\right]^{-T/2} h_{1,\mu}(\mu).$$

For details, see a standard textbook such as Zellner (1971).

A researcher will hold some belief about the veracity of each model and this belief is usually represented by the prior probabilities for each model, $\Pr(M_i)$. Given this prior belief we can obtain the posterior probabilities of two models from the posterior odds ratio which is defined by

$$\begin{aligned}\frac{\Pr(M_i|y)}{\Pr(M_j|y)} &= \frac{\Pr(M_i)}{\Pr(M_j)} \frac{c_i \int_\Theta f_i(y|\theta_i) h_i(\theta_i)}{c_j \int_\Theta f_j(y|\theta_j) h_j(\theta_j)} \\ &= \frac{\Pr(M_i)}{\Pr(M_j)} B_{ij}.\end{aligned}$$

The first term in the above expression, $\Pr(M_i)/\Pr(M_j)$, is the prior odds ratio and the second term, $B_{ij}$, is the Bayes factor. For the remainer of this paper we will assume equal prior probabilities for the models under comparison such that $\Pr(M_i)/\Pr(M_j) = 1$ and our interest is in calculating $B_{ij}$. A useful representation of the Bayes factor for model $M_i$ to model $M_j$ is as the ratio of the marginal likelihoods $m_i = p_i c_i$,

$$B_{ij} = \frac{m_i}{m_j} = \frac{p_i}{p_j}\frac{c_i}{c_j}. \tag{2}$$

For a range of reasons, including actual ignorance, it is common practice to endeavour to report results that reflect no or weak prior beliefs, ie. ignorance, about a model or its parameters. Therefore improper priors on the parameters and equal prior probabilities on the models are used to represent this uninformativeness or ignorance and only the resulting Bayes factors are reported rather than the posterior probabilities for two models of interest. A problem arises, however, in calculating Bayes factors when using improper priors on the parameters of interest.

As is clear from the expression (2), the Bayes factor is proportional to the ratio $\varsigma = c_i/c_j$ under the conditions that $p_i$ and $p_j$ are convergent integrals. If both priors are improper then the constants $c_i$ and $c_j$ are infinite as nonconvergent integrals such that the ratio $\varsigma$ is undefined. Usually this result arises because the dimensions of the parameter spaces differ, such as occurs when one model is nested within another. If only one of the priors is improper such that only $c_i$ or $c_j$ is infinite, then $B_{ij} = 0$ or $\infty$ *a priori*. This effect can be demonstrated for the zero mean example discussed earlier by beginning with a proper flat prior density for $\mu$ over $[-M, M]$ under the alternative. The prior and posterior densities for this example are presented in Figure 1. The Bayes factor for the zero mean model to the model with mean in the interval $[-M, M]$ has the form

$$B_{01} = \frac{\left[\nu s^2 + T\overline{x}^2\right]^{-T/2}}{\int_{-M}^{M}\left[\nu s^2 + T(\mu - \overline{x})^2\right]^{-T/2} d\mu \big/ M}.$$

For large $T$ and $M$, the integral $\int_{-M}^{M}\left[\nu s^2 + T(\mu - \overline{x})^2\right]^{-T/2} d\mu \approx \sqrt{\pi}\big/ (\nu s^2)^{(T-1)/2}$ such that $B_{01} \approx M\left[\nu s^2 + T\overline{x}^2\right]^{-T/2} (\nu s^2)^{(T-1)/2} \big/ \sqrt{\pi}$. As we increase $M$ towards $\infty$, we see $B_{01} \longrightarrow \infty$ and the null will be prefered regardless of the information in the data. More extensive discussion of this issue can be found in Zellner (1971), Berger and Perrichi (1996), O'Hagan (1995) and Lindley (1997).

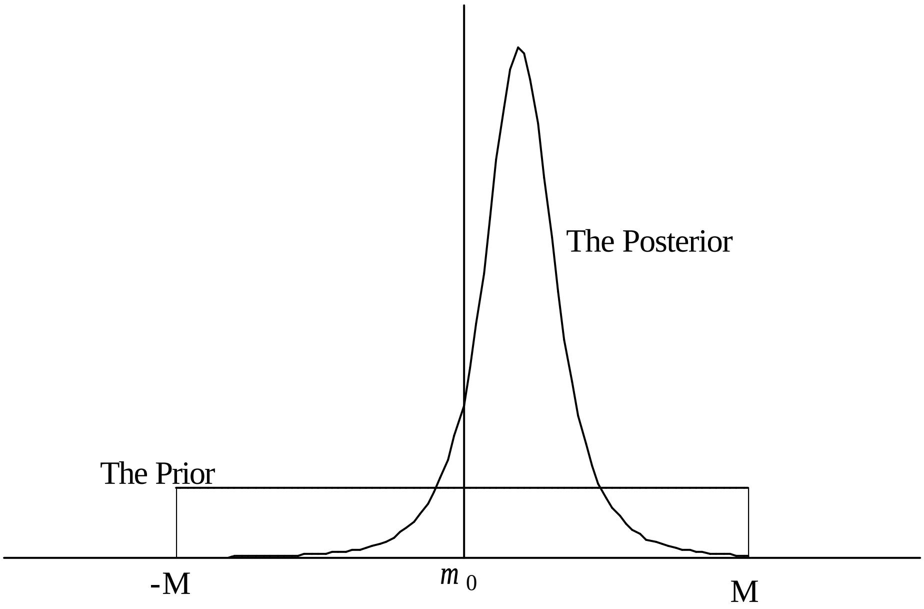

Figure 1: The sharp null prior occurs at $\mu_0$ and the diffuse alternative extends from $-M$ to $M$.

A number of approaches have been proposed to deal with this issue. One approach is to reject improper priors as nonsensible for model comparison, other approaches allow the approximate calculation of posterior probabilities while retaining ignorance priors. This latter approach generally involves approximations to the Bayes factor.

Berger and Perrichi (1996) proposed developing a posterior using a minimum size training sample such that this posterior will be proper. O'Hagan (1995) presents another approach which follows the same principle of using a fraction of the sample to remove indeterminancy of the Bayes factor. However, it is nontrivial to determine which fraction of the data to use as a training sample. Robustness of this data-based prior approach needs to be investigated. One example of this approach is the Bayesian unit root analysis of Schotman and van Dijk (1991) where the initial value of the series of observations is used to construct an informative prior.

Lindley (1997) argues that only proper priors should be employed to represent uncertainty and used model comparison as one motivating example. However, it is nontrivial to find informative priors for rather complex models. Further, one may have an interest in the information content in the likelihood compared with 'vague' or 'diffuse' prior information. So, we still have no general guidelines as to how this principle could be implemented consistently across models and it is likely that researchers will continue to wish to present results with uninformative or ignorance priors.

## 2.1 Cointegration analysis and improper priors

A simple generalisation of the linear regression model is the multivariate linear model, that is, a system of linear equations. Here we consider as a special case the cointegration model. In simple terms, cointegration analysis involves the study of systems of variables that are nonstationary, such as variables that are $I(1)$, with the aim of finding linear combinations of these nonstationary variables that are stationary, such that these combinations may be, for example, $I(0)$. In cointegration analysis the researcher has two aims: the first aim is to find the number of stationary combinations that exist; and the second aim is to find the actual combinations. The coefficients that form these linear combinations are collected into vectors called cointegrating vectors and the number of these vectors match the dimension of the space. The two aims are met by estimating what is called the cointegrating space since we first estimate the dimension of the cointegrating space and then estimate the direction of the space given its dimension by estimating the cointegrating vectors.

A particular space implies a particular linear combination of the variables entering the cointegrating relation. Consider, for example, the following process $x_t = (x_{1,t}, x_{2,t})'$ in which $s_t \sim I(1)$.

$$\begin{aligned} x_{1,t} &= as_t + \nu_{1,t} \\ x_{2,t} &= bs_t + \nu_{2,t}. \end{aligned}$$

This bivariate system has a one dimensional cointegrating space. Taking a linear combination of the above processes

$$\beta_1 x_{1,t} + \beta_2 x_{2,t} = \beta' x_t = (\beta_1 a + \beta_2 b) s_t + \varepsilon_t$$

where $\beta = (\beta_1, \beta_2)'$, $\varepsilon_t = \beta_1 \nu_{1,t} + \beta_2 \nu_{2,t}$, we are particularly interested in the case $\beta_1 a + \beta_2 b = 0$ or $\beta_2 = -\frac{a}{b}\beta_1$ such that $\beta x_t \sim I(0)$. Note that any vector proportional to $\beta$, such as $\beta\kappa$,will also solve the relation $\beta_1^{-1}\kappa^{-1}\kappa\beta_2 = \beta_1^{-1}\beta_2 = -ab^{-1}$. The vectors which solve this relation lie on the dashed line in Figure 2 and this line is the cointegrating space. Note that both vector A and the negative of vector A will both lie in the cointegrating space. We present two examples of such vectors of different length in Figure 2- vectors A and B - to emphasise the point that the length of the vector is not important, or even its exact direction, only the direction of the space it spans (the dashed line in Figure 2). The quantity of interest might be thought of as $\tan\theta$ where $\theta$ is the angle shown in Figure 2. Thus if we were to restrict the vectors to unit length, that is use the Euclidean norm $\beta'\beta = 1$, this has no implications for inference on the cointegrating space.

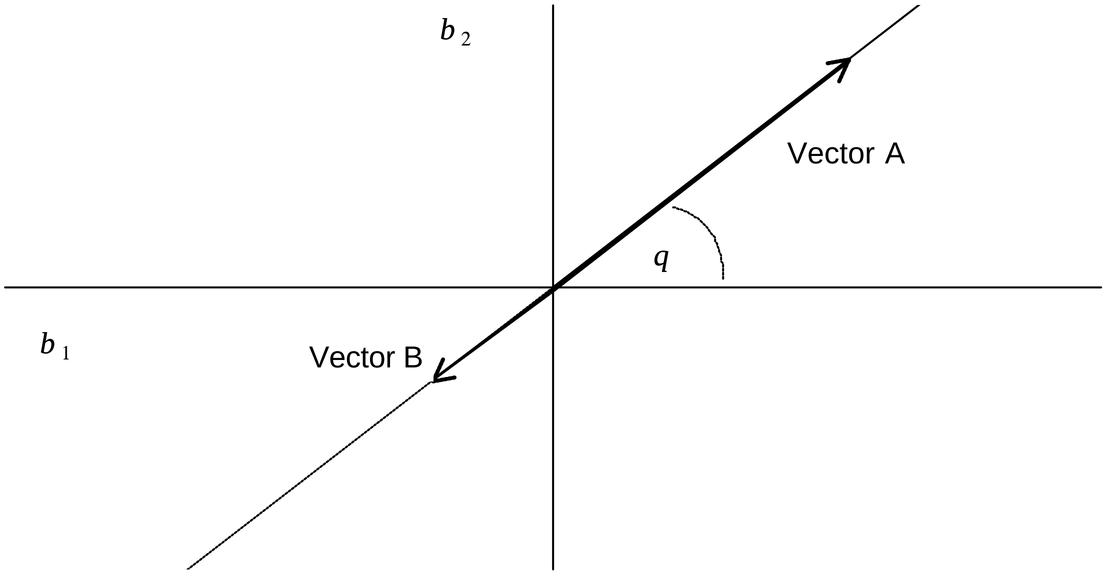

Figure 2: Although they are of different lengths and point in opposite directions, Vectors A and B span the same space (the dashed line).

We can generalise this visual analysis slightly to a trivariate system such as the following process $x_t = (x_{1,t}, x_{2,t}, x_{3,t})'$ in which the scalar $s_t \sim I(1)$ :

$$\begin{aligned} x_{1,t} &= as_t + \nu_{1,t} \\ x_{2,t} &= bs_t + \nu_{2,t} \\ x_{3,t} &= cs_t + \nu_{3,t}. \end{aligned}$$

This trivariate system has a two dimensional cointegrating space. Taking a linear combination of the above processes

$$\begin{aligned}\beta_{i1}x_{1,t} + \beta_{i2}x_{2,t} + \beta_{i3}x_{3,t} &= \beta' x_t \\ &= (\beta_{i1}a + \beta_{i2}b + \beta_{i3}c)\, s_t + \varepsilon_t\end{aligned}$$

where $\beta_i = (\beta_{i1}, \beta_{i2}, \beta_{i3})'$, $\varepsilon_t = \beta_{i1}\nu_{1,t} + \beta_{i2}\nu_{2,t} + \beta_{i3}\nu_{3,t}$. If $\beta_i$ is a cointegrating vector then $\beta_{i1}a + \beta_{i2}b + \beta_{i3}c = 0$. Assuming $\beta_{i1} = \delta\beta_{i2}$ does not hold for both $i = 1$ and $i = 2$, it can be shown that all solutions to this restriction will be some linear combination of the vectors

$$\begin{bmatrix} 1 \\ 0 \\ -\frac{a}{c} \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 1 \\ -\frac{b}{c} \end{bmatrix}.$$

An example of such two vectors are shown in Figure 3. It can be seen that the vectors lie on a two dimensional plane (the cross-hatched plane in Figure 3) and any linear combination of these vectors will lie in this plane and this plane is what we refer to as the cointegrating space. So to restate more clearly our interest in the cointegrating vectors - or our objective in estimating the cointegrating space - we are not interested in the length of the vectors, nor in their direction specifically. Our interest is in the direction of the space spanned by by the vectors.

A distinguishing feature of these combinations is that there is no dependent variable *per se* in a cointegrating relation upon which one could sensibly normalise in a regression. However some form of normalisation is required to identify the elements of the cointegrating vectors.

Linear identifying restrictions - in which the relations between a set of coefficients is assumed known - is almost always employed in Bayesian cointegration analysis. Imposing linear identifying restrictions has several limitations of which we mention the following ones. First it assumes that we know something about the cointegrating relations which we may rather not assume and, as demonstrated in Strachan (2003), may turn out to be invalid. Even if one has correctly chosen the coefficients upon which to normalise there remain several theoretical and practical issue with obtaining inference with linear restrictions. For example, the posterior for the cointegrating vectors which have had linear identifying restrictions imposed, has been shown in several studies to be of a rather peculiar form (see, for example, Bauwens and Lubrano 1996 and Kleibergen and van Dijk 1994). This brings us to the

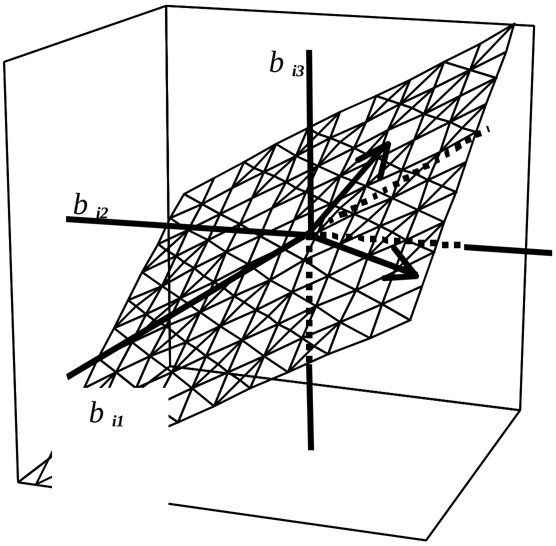

Figure 3:

second limitation that the general features of the posterior in this case are that it is bimodal and has no moments. As inference on the cointegrating space proceeds from the estimates of the cointegrating vectors, this inference may be difficult to obtain if these cointegrating vectors have no mean to estimate and a unique, global, mode is difficult to obtain. Marginal medians offer a possible alternative, however, again, obtaining these even in simple models is not as simple or efficient as estimating an extant mean. For the purpose of model averaging, obtaining expectations from each model to average over the models is a more consistent procedure than averaging modes or medians.

When employing linear identifying restrictions, a third serious issue arises in particular specifications of cointegrating models. Kleibergen and van Dijk (1994, 1998) demonstrate a seemingly sensible specification in which the resulting posterior for the cointegrating vectors is not proper. Another important model feature which is commonly employed in cointegration analysis and has a strong economic justification in many cases, is weak exogeneity. This restriction is commonly investigated and imposed as it has either or both theoretical as well as empirical support in many applications. However, as demonstrated in Strachan and van Dijk (2003), a fourth limitation of linear identifying restrictions is that when coupled with weak exogeneity an improper posterior results. It is difficult to imagine a more serious impediment to Bayesian inference than an improper posterior.

# 3 Orthogonal regression and proper diffuse priors.

Given the problems associated with employing linear identifying restrictions with diffuse priors in Bayesian cointegration analysis, it makes sense to estimate the cointegrating vectors prior to normalisation on a particular coefficient. This can be achieved using Euclidean normalisation which is, in very special cases, equivalent to orthogonal regression and in general displays the same features as orthogonal regression. In this paper we demonstrate the application in a form equivalent to orthogonal regression. Using Euclidean normalisation offers a way of obtaining posterior probabilities, posterior moments, and posterior modes even when such restrictions as weak exogeneity are imposed. Most importantly for model averaging, it allows us to use uninformative priors with well defined, finite normalising constants such that Bayes factors are well defined.

Reconsider the model for the mean of a series as given in (1) but respecify the model as

$$\alpha_0 x_t - \alpha_1 = \varepsilon_t = \alpha' y_t \tag{3}$$

where $\alpha = (\alpha_0, \alpha_1)'$, $y_t = (x_t, 1)'$ and $\alpha'\alpha = 1$. We ignore the rescaling of $\sigma^2$ from this transformation as this is easily dealt with in the Bayesian set up.

The specification in (3) can be linked to the one in (1) by the obvious transformation $\mu = \alpha_1/\alpha_0$. Although there is a simple relationship between normal regression and orthogonal regression, there are implications for inference from this respecification. One example is for the interpretation of the least squares method.

In least squares estimation, for example, the error is taken to be a linear combination of $x_t$ and 1 with a normalisation on the coefficient for $x_t, \varepsilon_t = x_t - \mu$. Thus the estimate of $\mu$ is obtained by minimising (the quadratic mean of) this 'vertical' distance between $x_t$ and the estimated line (or plane) $\widehat{\mu}$. This distance is shown in Figure 4 as $\varepsilon_A$ and we minimise , $\Sigma_{t=1}^{T} \varepsilon_{tA}^2$. If we were instead to estimate from the regression of 1 on $x_t$, $\varepsilon_t = 1 - \gamma x_t$, and take $\widehat{\gamma}^{-1}$ as our estimate of $\mu$, the estimator would minimise the 'horizontal' distance

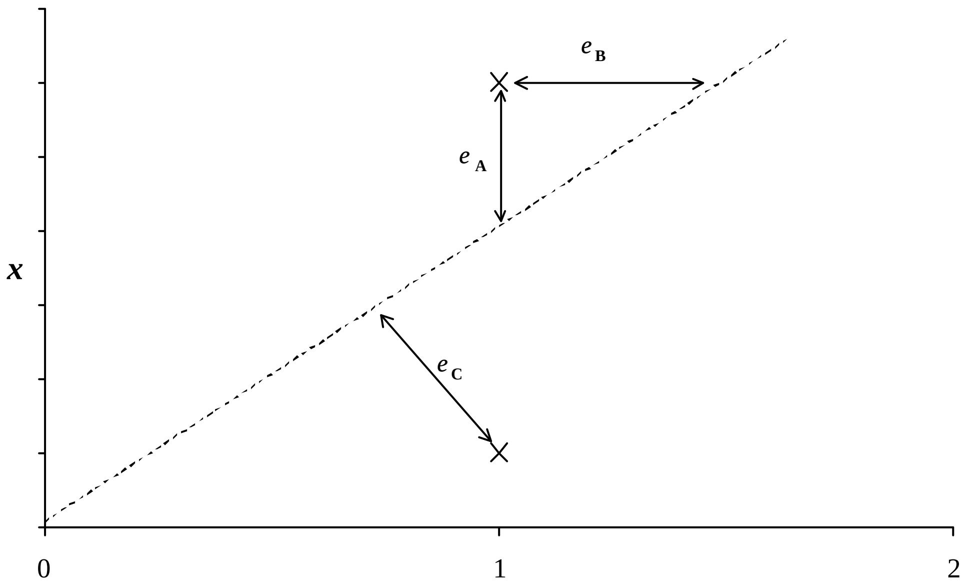

Figure 4:

between 1 and the estimated line (or plane) $\widehat{\gamma}x_t$, the distance denoted in Figure 4 as $\varepsilon_B$. In this case we minimise $\Sigma_{t=1}^T\varepsilon_{tB}^2$. However, in orthogonal regression the distance that is minimised is the shortest distance between the point $(x_t, 1)$ and the line implied by $\widehat{\alpha}y_t = 0$, $\varepsilon_t = \alpha_0 x_t - \alpha_1 = \alpha' y_t$ (Malinvaud, 1970), the distance denoted as $\varepsilon_C$ in Figure 4. We minimise $\Sigma_{t=1}^T\varepsilon_{tC}^2 = \Sigma_{t=1}^T(\alpha' y_t)^2$ subject to $\alpha'\alpha = 1$ and take $\widehat{\alpha}_1/\widehat{\alpha}_0$ as our estimate of $\mu$. This approach defines orthogonal least squares regression as an eigenvalue-eigenvector problem.

Next we consider the form of the marginal posterior density for $\alpha$ with the restriction $\alpha'\alpha = 1$. Taking the set of parameters in the model as $\theta_0 = (\sigma^2, \alpha)$, then given a prior $h_0(\theta_0)$, the posterior is now proportional to

$$f_0(y|\theta_0)\,h_0(\theta_0) = \sigma^{-T}\exp\left\{-\frac{1}{2\sigma^2}\alpha'\left(\Sigma_{t=1}^T y_t y_t'\right)\alpha\right\}h_0(\theta_0)\,p_0^{-1}. \qquad (4)$$

Again apply the prior for $\sigma^2$ of $\sigma^{-1}$ such that $h_0(\theta_0) = h_{0,\alpha}(\alpha)/\sigma$. Integrating with respect to $\sigma^2$ we obtain the posterior for $\alpha$ *proportional* to

$$\left|\alpha'\left(\Sigma_{t=1}^T y_t y_t'\right)\alpha\right|^{-T/2} h_0(\alpha). \qquad (5)$$

To impose the restriction to model $M_1$ implied by $\mu = 0$, we set $\alpha_1 = 0$ and therefore $\alpha_0 = 1$. The prior is simply a discrete prior probability mass function $\pi_1(\alpha_0) = 1$ for $\alpha_0 = 1$ and zero otherwise. The resultant marginal posterior (marginalised with respect to $\sigma^2$) is then $\left(\Sigma_{t=1}^T x_t^2\right)^{-T/2}$.

In the following subsection we complete the posterior in (5) by developing the explicit form for $h_0(\alpha)$.

## 3.1 A proper diffuse prior

An important implication of this specification for the estimation of Bayes factors and posterior probabilities is that the support for the parameters is now compact. This implies that the normalising constant for the prior is known such that the Bayes factors for comparing models of different dimensions will be well defined. That is the restriction $\alpha'\alpha = 1$ implies $\alpha_1 \in [-1, 1]$ and $\alpha_0 = \sqrt{1 - \alpha_1^2}$ and the support for the vector $\alpha$ describes the upper hemishere of a unit circle centered at the origin. We only care about the upper half hemisphere since we are interested in the implied value for $\mu$, and the two vectors $\alpha$ and $-\alpha$ imply the same value for $\mu$ since $\mu = \alpha_1/\alpha_0 = (-\alpha_1)/(-\alpha_0)$. The vectors $\alpha$ and $-\alpha$ lie on the same line passing through the origin, that is they span the same space. Thus we only care about the space spanned by the vector $\alpha$.

A useful representation of the vector $\alpha$ to demonstrate this concept is in polar coordinates. That is we express the direction of $\alpha$ as a function of a single angle, $\theta \in [0, \pi]$. Thus, $\alpha_0 = \cos\theta$ and $\alpha_1 = \sin\theta$ such that $\mu = \tan\theta$. The parameter of interest no longer refers to the particular coefficients, rather it is the direction (or space) of the vector $\alpha$ or, equivalently, the angle $\theta$, and we can think of the prior on the direction as a prior on $\theta$. If we have no prior beliefs about the direction for the vector $\alpha$, we can represent this with a uniform prior on the support for $\theta$. Thus we have the prior $\pi_0(\theta) = c_0 = \pi^{-1}$.

Using angles is useful for exposition in the simple bivariate case, however it is not simple to undertake analysis using angles as parameters of interest in the general case. Moreover, this approach produces a Jacobian that grows with the dimension of the model and involves more accounting of parameters than simply regarding the vector (or, in some applications, the matrix) as the parameter of interest.

Next, we are going to make use of some results on manifolds, in particular the notion of a Stiefel manifold. The support for the $n-$dimensional vector $\alpha$

subject to $\alpha'\alpha = 1$ is known as a the Stiefel manifold and denoted as $V_{1,n}$. We can think of this space as an $n-$dimensional sphere centered at the origin. To derive a uniform distribution for $V_{1,n}$, we can use the fact that the direction of an $n-$dimensional vector $z$ in which each element of $z$ is independently distributed as standard normal is uniformly distributed over $V_{1,n}$ (James, 1954). Since $x$ has Euclidean length $l = (z'z)^{\frac{1}{2}}$ , we can decompose $z$ into $l$ and $\alpha = zl^{-1}$, integrate with respect to $l$ to obtain the uniform prior for $\alpha$ on $V_{1,n}$. As we are only interested in the half sphere, we take as our normalising constant

$$\pi_0(\alpha) = \frac{\Gamma\left(\frac{n}{2}\right)}{\pi^{\frac{n}{2}}} = c_0.$$

To link this idea with the earlier discussion about the angle of the vector, $\theta$, we note that in the bivariate case $(n = 2)$ discussed above this flat prior on $V_{1,2}$ implies a flat prior on $\theta$.

Jeffreys (1961) argued for the use of proper priors for model selection and hypothesis testing. One prior he argued for was the Cauchy prior for the parameter of interest which is consistent with our prior. Partition $\alpha$ as $\alpha = (\alpha_0, \alpha_1)$ where $\alpha_0$ is the first element of $\alpha$ such that $\alpha_1$ is the vector of the remaining elements. The Jacobian for the transformation from $\alpha$ to $\mu = \alpha_1/\alpha_0$ is proportional to $(1 + \mu'\mu)^{-n/2}$ (see Phillips 1994, Appendix p. 86) which is the kernel for the Cauchy density. That is, a flat prior on the direction of the vector implies a Cauchy prior for $\mu$. This last implication is a general result for the transformation from the Stiefel manifold to the real line.

We now briefly discuss some of the important properties of the posterior. Using the flat prior on $V_{1,n}$, the resulting posterior in (5) is proportional to

$$k(\alpha) = \left|\alpha'\left(\Sigma_{t=1}^{T} y_t y_t'\right)\alpha\right|^{-T/2}.$$

Denote the smallest eigenvalue of the matrix $\left(\Sigma_{t=1}^{T} y_t y_t'\right)$ by $\lambda$. As $\lambda$ will be positive with probability one, we know by the Poincaré separation theorem (see, for example, Schott 1997) that the function $\alpha'\left(\Sigma_{t=1}^{T} y_t y_t'\right)\alpha$ subject to $\alpha'\alpha = 1$ will have a finite lower bound $\lambda$. This implies that the function $k(\alpha)$ will have the finite upper bound $v = \lambda^{-T/2}$. As the support for $\alpha$ is compact, we know the integral $\Upsilon = v\int(\alpha' d\alpha)$ will be finite. As the integral $\int k(\alpha)(\alpha' d\alpha)$ will be less than $\Upsilon$, we can therefore say that the posterior will be proper and all finite absolute moments will exist. This is important for model averaging as we can now say that $p_i$ is finite, such that $B_{ij}$ is well defined as are the posterior probabilities.

As we know that the posterior is proper, that is $p_i$ is finite, and we know the prior normalising constant $c_i$, from expression (2) we can see that our Bayes factors and therefore our posterior probabilities, will be well defined. In this model above, the Bayes factor will have the form $B_{ij} = \frac{m_i}{m_j} = \frac{p_i c_i}{p_j c_j}$ where

$$p_i = \int_{V_{1,n}} \left| \alpha' \left( \Sigma_{t=1}^{T} y_t y_t' \right) \alpha \right|^{-T/2} (\alpha' d\alpha)$$

and

$$c_i^{-1} = \int_{V_{1,n}} (\alpha' d\alpha) .$$

The integral for $c_i$ is known and an example has been given above. However, as the integral in $p_i$ is often of a very nonstandard form, and analytical expressions for the integral do not have a closed form (this can be shown using the expression given in Muirhead 1982, p. 279) except in simple cases, it is usual to use an approximation. Two approximations that have been used are the Laplace approximation (James 1969 and Strachan and Inder 2003) and Markov Chain Monto Carlo integration (Strachan 2003).

A comment on the notation for the differential term $(\alpha' d\alpha)$ is in order. This measure is in the exterior product differential form. Interested readers are directed to Muirhead (1982) and James (1954) for details on this topic. For our purposes, we can think of $(\alpha' d\alpha)$ as doing the job of the usual differential for $x$, $(dx)$, the only difference is that $\alpha$ has a compact support which forms a manifold in $\mathbb{R}^n$.

# 4 Posterior probabilities of alternative models.

To demonstrate the implementation of this approach for the evaluation of posterior probabilities we outline some simple examples. We begin with the well known and well understood simple and multiple linear regression models, and then present a simple model of a cointegrated system. This model of cointegration slightly generalises that used by Phillips (1994), but is more restrictive than the vector error correction model. However, it is sufficient to demonstrate the main features of the approach.

### 4.0.1 Estimating regression equations as vector spaces

In this example we take the simple linear regression model which one would estimate if it were thought that a linear, or near linear relationship is most appropriate for the variables under investigation. Thus we have the standard form for the linear regression model as

$$x_{0,i} = \beta_0 + \beta_1 x_{1,i} + \nu_i.$$

Using our specification of orthogonal regression we have

$$\begin{aligned} \alpha_0 x_{0,i} + \alpha_1 + \alpha_2 x_{1,i} &= \varepsilon_i \\ \alpha' y_i &= \varepsilon_i \end{aligned}$$

where $\alpha' = (\alpha_0, \alpha_1, \alpha_2)$, $y_i = (x_{0,i}, 1, x_{1,i})'$ and $\alpha'\alpha = 1$.

For the multiple regression model in which $x_{1,i}$ and $\alpha_2'$ are $(n-2) \times 1$ vectors and so $\alpha$ is an $n \times 1$ vector and $\beta = (\beta_0, \beta_1, \ldots, \beta_{n-2})$, $y_i = (x_{0,i}, 1, x_{1,i}')'$ and $\alpha'\alpha = 1$. Assuming the prior in the previous section we obtain the posterior

$$\pi_1(\alpha|y)(\alpha' d\alpha) = c_1 p_1^{-1} \left|\alpha' \left(\Sigma_{t=1}^T y_t y_t'\right) \alpha\right|^{-T/2} (\alpha' d\alpha).$$

We wish to calculate the posterior probability that $\alpha_0 = \alpha_2$ (implying $\beta_1 = \frac{\alpha_2}{\alpha_0} = 1$) using an uninformative prior for $\alpha$. The form for the posterior is then achieved by replacing $\alpha$ by $R\widetilde{\alpha} = (\alpha_0, \alpha_1, \alpha_0, \alpha_3, \ldots, \alpha_{n-1})$ and we have $\widetilde{\alpha}'\widetilde{\alpha} = 1$. The $n \times (n-1)$ matrix $R$ is defined by its orthogonal compliment $R_\perp = (1, 0, -1, 0, \ldots, 0)'$ such that $R_\perp' R = 0$. The posterior for this restricted model is then

$$\pi_0(\widetilde{\alpha}|y)(\widetilde{\alpha}' d\widetilde{\alpha}) = c_0 p_0^{-1} \left|\widetilde{\alpha}' R' \left(\Sigma_{t=1}^T y_t y_t'\right) R\widetilde{\alpha}\right|^{-T/2} (\widetilde{\alpha}' d\widetilde{\alpha}).$$

As an example we consider the regression of the log excess return for an asset $(r_t)$ upon the log excess return of the market portfolio $(m_t)$ to which this asset belongs. We use data on the 30 day bank bill rate, the price for the National Australia Bank ordinary shares and the Australian All Ordinary Shares Index which is the standard index for the general level of the Australian stock exchange. The data are monthly (end of month) figures covering the period from December 1988 to June 1988 for a total of 139 observations. The models we are interested in comparing are $r_t = \beta m_t + \varepsilon_t$

with $\beta = 1$ $(M_0)$ and $\beta \neq 1$ $(M_1)$. We use monthly figures to aggregate out much of the volatility clustering. This seems successful and we note that when we estimate this model, although the individual series $r_t$ displays some evidence of correlation in the squared residuals, our estimate of $\varepsilon_t$ does not. The least squares estimate of $\beta$ is 0.7068 while the orthogonal regression estimate is 0.9714, very close to the $M_0$ value. However, we are still able to distinguish this estimate from the value of one implied by $M_0$. When we compute the posterior probabilities of $M_0$ and $M_1$ we find unambiguous support for $M_1$ with $\Pr(\beta \neq 1|y) = 1$.

| Estimates for $\beta$ | | | |
|---|---|---|---|
| OLS | Orthogonal Regression | $\Pr(\beta = 1\|y)$ | $\Pr(\beta \neq 1\|y)$ |
| 0.707 | 0.971 | 0 | 1.0 |

### 4.0.2 Cointegration

Our interest in this subsection is in estimating the dimension and space spanned by the cointegrating space. This is probably the most natural setting for the approach we propose as in cointegrating relations there is no clear 'dependent' variable upon which to normalise. As demonstrated in related work (Strachan 2003, Strachan and Inder 2003 and Strachan and van Dijk 2003) normalising can in fact complicate attempts to obtain inference. If we wish to normalise and thereby effectively select a dependent variable, taking the orthogonal regression approach allows us to evaluate our selection, however we do not go so far as to investige how to do this in this paper. Our objective in this section is to demonstrate estimation of the dimension of the space and the space itself.

We 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 spread between rates of different maturity are $I(0)$ and thus form a cointegrating relation. The data for this example is 94 observations of the 5 year $(i_5)$ and 3 year $(i_3)$ Australian Treasury Bond rates from July 1992 to April 2000. A plot of these series is presented in Figure 5 along with the spread, $s_t = i_{5,t} - i_{10,t}$. From this plot we see the unstable nature of $i_{5,t}$ and $i_{10,t}$ but the relatively stable behaviour of $s_t$.We will employ an error correction model to demonstrate the important aspects of this method in cointegration analysis. This model nests the model used by Phillips (1994) to investigate the finite sample behaviour of his estimator for the cointegrating vectors and has the form

$$\Delta x_t = \mu + \alpha\beta' x_{t-1} + \varepsilon_t$$

where $x_t = (i_{5,t}, i_{10,t})$ and $\varepsilon_t \sim i.i.d.N\left(0, \Sigma\right)$. We lose one observation due to the lag term. The structure of this model appears sufficient in the sense that both Portmanteau and LM tests for autocorrelation accept the null of no autocorrelation in the residuals. Although this is a restrictive model in some senses, it is sufficient to demonstrate the main features of this technique.

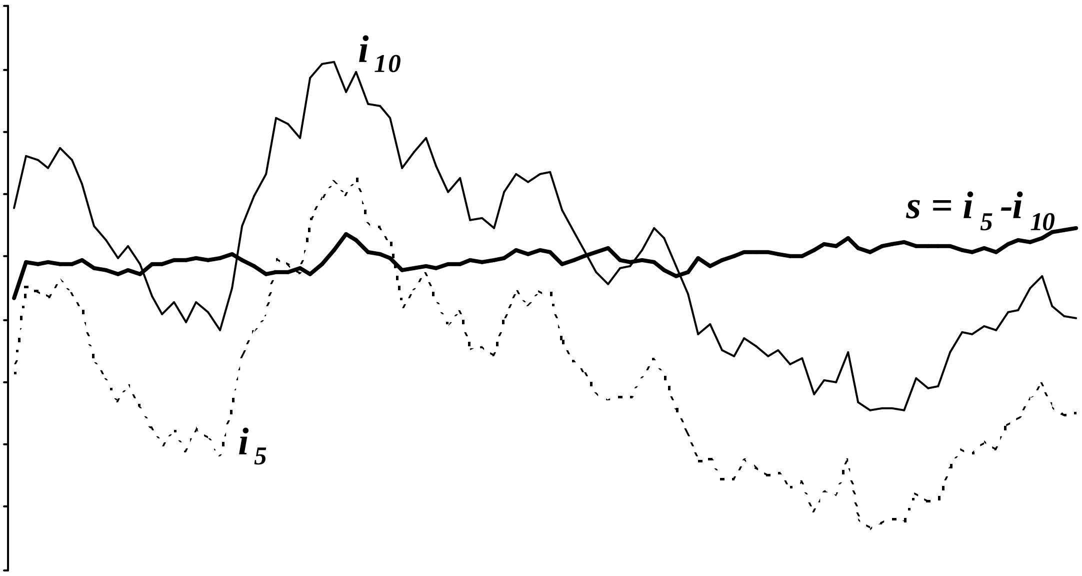

Figure 5: This figure shows the plot of $i_5$, $i_{10}$ and the spread, $i_5 - i_{10}$, over the period July 1992 to April 2000. The level of each series has been adjusted to clarify the dynamic relationships among the variables.

Define $\beta$ to be semiorthogonal such that $\beta'\beta = I_r$. We have previously described the Stiefel manifold $V_{1,n}$ as the space spanned by an *n-vector* of unit length and presented the volume of this space. As the restriction $\alpha'\alpha = 1$ imposes one restriction on the vector, the space $V_{1,n}$ is an $n - 1$ dimensional manifold in $n$-space. Next we describe the space spanned by the semiorthogonal *matrix* $\beta$. Each vector in $\beta$ is of unit length, but the vectors are all orthogonal to eachother and the restriction $\beta'\beta = I_r$ imposes $\frac{r(r+1)}{2}$ restrictions such that the parameter space is $nr - \frac{r(r+1)}{2}$ dimensional. Thus each vector in $\beta$ describes an $n$-dimensional sphere centered at the origin. This decribes an element of the Stiefel manifold, $V_{r,n}$, such that we can say $\beta \in V_{r,n}$.

Placing a flat prior on $V_{r,n}$ such that we have an uninformative prior for $\beta$, results in the prior[1]

$$\pi(\beta)(\beta' d\beta) = c_r h(\beta)(\beta' d\beta)$$

where

$$\begin{aligned} h(\beta) &= 1, \text{ and} \\ c_r^{-1} &= \int_{v_{r,n}} (\beta' d\beta) \\ &= \frac{2^r \pi^{nr/2}}{\Gamma_r\left[\frac{n}{2}\right]}, \text{ where} \\ \Gamma_r\left[\frac{n}{2}\right] &= \pi^{r(r-1)/4} \Pi_{i=1}^r \Gamma\left[(n+1-i)/2\right]. \end{aligned}$$

Again we direct the reader to Muirhead (1982) for further discussion on the differential form $(\beta' d\beta)$.

To reiterate our aim, we wish to estimate $r$ and the space spanned by $\beta$. There has been considerable work in the literature on Bayesian cointegration analysis outlining the issues associated with local nonideintification (see for example Kleibergen and van Dijk 1994, Martin & Martin 2000, Martin 2000, and Martin 2001). It is worth mentioning that our approach will not result in the same problems associated with local nonidentification (see Strachan and van Dijk 2003).

The likelihood can be expressed as

$$L(\mu, \alpha, \beta, \Sigma) = (2\pi)^{-Tr/2} |\Sigma|^{-T/2} \exp\left\{-\frac{1}{2} tr \Sigma \left(\Sigma_{t=1}^T \varepsilon_t \varepsilon_t'\right)\right\}.$$

Using a standard prior for $\Sigma$ of $|\Sigma|^{-3/2}$ and combining the priors with the likelihood gives the posterior proportional to

$$c_r (2\pi)^{-Tr/2} |\Sigma|^{-(T+r+1)/2} \exp\left\{-\frac{1}{2} tr \Sigma \left(\Sigma_{t=1}^T \varepsilon_t \varepsilon_t'\right)\right\}.$$

[1] This prior does not take into account that the parameter of interest is a Grassman manifold which is of a lower dimension than the Stiefel manifold. This issue is discussed fully in related work (see Strachan and van Dijk, 2003).

As is shown in Strachan and van Dijk (2003), integrating with respect to $\Sigma$, $\mu$ and $\alpha$ we obtain the marginal posterior distribution for $(\beta, r)$ as proportional to

$$2^{-r}\pi^{-Tr/2}\Pi_{i=1}^{r}\frac{\Gamma\left[(T+1-i)/2\right]}{\Gamma\left[(3-i)/2\right]}\left|\beta' D_0 \beta\right|^{-T/2}\left|\beta' D_1 \beta\right|^{(T-2)/2}.$$

A range of approaches now exist to approximate the integral with respect to $\beta$ on the Stiefel manifold. Papers presenting Monte Carlo integration over this space include Strachan (2003) and Strachan and van Dijk (2003). Strachan and Inder (2003) present an approach using the Laplace approximation to the same integral. In this paper we use a first order asymptotic approximation similar to that presented in Schwarz (1978), but we incorporate the correct normalisation of the integral for the cointegrating space.

The cointegrating relation may be written as $\beta' x_t = b_1 i_5 - b_2 i_{10}$ and for the hypothesis that the spread is stationary (while $x_t \sim I(1)$) we are interested in estmating $b = b_2/b_1$. Using the maximum likelihood estimation procedure to estimate $b_1$ and $b_2$ we obtain an estimate of $b$ equal to -0.9454. The Bayesian estimate of $b$ is -0.9386. The classical trace and maximum likelihood tests for the rank select $r = 1$. The estimated Bayesian posterior probability of the different values for $r$ are presented in the table below in the second column. While there is some support for $r = 1$ ($\Pr(r = 1|y) = 0.24$), the mass of the support is for $r = 2$.

| Models: $M_i$ | $\Pr(M_i\|y)$ | $\Pr(M_i\|y)$ |
|---|---|---|
| $M_0 : r = 0$ | 0.004 | 0.002 |
| $M_1 : r = 1$ | 0.240 | 0.130 |
| $M_2 : r = 1$ and $\beta = (1, -1)'$ | - | 0.457 |
| $M_3 : r = 2$ | 0.756 | 0.411 |

Next we expand the model set to include the model in which the spread is treated as the cointegrating relation, that is, the model with $r = 1$ and $\beta = (1, -1)'$. Classical likelihood ratio test for this restriction is not rejected at the 5% level of significance ($p$-value = 0.065). The estimated posterior probabilities for these models are presented in the third column of the table above. This additional model is the most probable model in the set with a probability of 45.7%. The support for $r = 1$ is now quite strong with posterior probability of 58.7% while the relative support for $r = 2$ falls now to 41.1%.

It is of interest to observe that a classical analysis would give all weight to a particular model ($r = 1$) but that a Bayesian approach gives nonnegligible weights to models with $r = 1$ and $r = 2$. Thus, indicating that while there is evidence that this feature of the term structure argument of Campbell and Shiller (1987) holds, there is also evidence that both series may by stationary. One may use these results for a forecasting exercise with model averaging.

# 5 Final Remarks.

In this paper we proposed a potential solution to the conflict in Bayesian inference between the desire to employ diffuse priors to represent ignorance and the desire to report proper posterior probabilities to alternative models. Using the concept of Stiefel manifolds we specified diffuse priors on dimension and direction of subspaces of parameter spaces within the context of a linear regression model and a cointegration model. We illustrated the approach using a CAPM and a term structure of interest rates model.

The proposed approach may be extended in several ways. Here, we mention the following ones. In financial econometrics one is often interested in the number of factors of an Arbitrage Pricing Theory (APT) factor model, see e.g. Campbell, Lo and MacKinlay (1997, Chapter 6). Issues of endogeneity and overidentification within systems of equations can also be analysed.

An other extension is to consider the issue of cointegration and the number of unit roots within the context of models that have a richer dynamics, more deterministic terms and possibly nonlinear characteristics such as Markov-switching. We note that the extension to lagged variables is not trivial.

Finally, within the class of large and empirically richer models one faces the issue of efficient computation of the posterior probabilities. Markov Chain Monte Carlo and Laplace approximations may be used. We refer to Strachan and Van Dijk (2003) for a study where a medium size model involving ten stochastic equations is successfully analysed.

# 6 References.

Bauwens, L. and M. Lubrano, 1996, Identification restrictions and posterior densities in cointegrated Gaussian VAR systems, in Advances in Econometrics, Vol. 11B, *Bayesian Methods Applied to Time Series Data*, T.B. Fomby, ed., (JAI Press) 3-28.

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

Campbell J. Y., A. W. Lo and A. C. MacKinlay (1997), *The Econometrics of Financial Markets,* Princeton University Press, Princeton, New Jersey.

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

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.

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

Jeffreys, H. (1961), *Theory of Probability,* London: Oxford University Press.

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.

O'Hagan A. (1995), "Fractional Bayes factors for model comparisons" *Journal of the Royal Statistical Society,* Series B, 57, 99-138.

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

Malinvaud, E., 1970, Statistical Methods of Econometrics, North-Holland, Amsterdam.

Martin, G.M., 2000, US Deficit Sustainability: a New Approach Based on Multiple Endogenous Breaks, Journal of Applied Econometrics, 15, 83-105.

Martin, G.M., 2001, Bayesian Analysis of a Fractional Cointegration Model, Econometric Reviews, Vol. 20, No. 2, 217-234 .

Martin, G.M.and V.L. Martin, 2000, Bayesian Inference in the Triangular Cointegration Model Using a Jeffreys Prior, Communications in Statistics, Theory and Methods, 29, No. 8.,1759-1785.

Muirhead, R. J., 1982, *Aspects of Multivariate Statistical Theory,* John Wiley and Sons, New York.

Phillips, P.C.B., 1994, “Some Exact Distribution Theory for Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models”, *Econometrica*, 62, 1, 73-93.

Schotman, P. and H. K. van Dijk, (1991), “A Bayesian analysis of the unit root in real exchange rates”, *Journal of Econometrics*, 49, 195-238.

Schott, J.R. (1997). *Matrix Analysis for Statistics.* John Wiley and Sons, New York.

Schwarz, G., (1978), Estimating the dimension of a model, *Annals of Statistics*, 6:2, 461-464.

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

Strachan, R. W. and B. Inder, 2003, “Bayesian Analysis of The Error Correction Model.”, *Journal of Econometrics*, forthcoming.

Strachan, R.W., and van Dijk, H. K. (2003), “The value of structural information in the VAR”, Discussion paper, University of Liverpool, Liverpool.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics,* John Wiley and Sons Inc., New York.

Zellner, A. (1986) “On assessing prior distributions and Bayesian regression analysis with $g$-prior distributions”, in: *Bayesian Inference and Decision Techniques - Essays in Honour of Bruno de Finetti,* eds., P.K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233-423.