Central European Journal of Economic Modelling and Econometrics

Divergent Priors and Well Behaved Bayes Factors

Rodney W. Strachan*, Herman K. van Dijk†

Submitted: 20.02.2014, Accepted: 22.03.2014

Abstract

Bartlett's paradox has been taken to imply that using improper priors results in Bayes factors that are not well defined, preventing model comparison in this case. We use well understood principles underlying what is already common practice, to demonstrate that this implication is not true for some improper priors, such as the Shrinkage prior due to Stein (1956). While this result would appear to expand the class of priors that may be used for computing posterior odds, we warn against the straightforward use of these priors. Highlighting the role of the prior measure in the behaviour of Bayes factors, we demonstrate pathologies in the prior measures for these improper priors. Using this discussion, we then propose a method of employing such priors by setting rules on the rate of diffusion of prior certainty.

Keywords: improper prior, Bayes factor, marginal likelihood, shrinkage prior, measure

JEL Classification: C11, C52, C15

*School of Economics, University of Queensland, Australia; email: r.strachan@uq.edu.au
†Econometric Institute, Erasmus University Rotterdam; email: hkvandijk@few.eur.nl

1 Introduction

This paper has three aims. First, we establish classes of priors that are exceptions to Bartlett's paradox. Second, we demonstrate pathologies in the Bayes factors that result from using improper priors. Finally, we present a method of obtaining well defined and well behaved Bayes factors with these classes of priors by controlling the rate of diffusion of certainty.

In empirical economic analysis, a natural extension of the concern for uncertainty associated with stochastic variables and parameter estimators is concern for uncertainty associated with the statistical or economic model used. While a common approach to data analysis is to select the 'best' of a set of competing models and then condition upon that model, this ignores the uncertainty associated with that model. An attractive feature of the Bayesian approach is the natural way in which model uncertainty may be assessed and incorporated into the analysis via the posterior model probabilities. An example of a method of incorporating this uncertainty that has attracted much attention in recent years is Bayesian model averaging (BMA). The benefits of BMA for prediction, for example, are outlined in several papers such as Min and Zellner (1993), Raftery, Madigan and Hoeting (1997) and Bernardo (1979). Another attractive feature of Bayesian analysis is the ability to incorporate the prior distribution $\pi(\theta) = h(\theta)/\mathfrak{c}$ where $\mathfrak{c} = \int h(\theta)\, d\theta$ is the unnormalised prior measure for the parameter space. This allows the researcher to reflect in the analysis a range of prior beliefs - from ignorance to dogma - that may reflect personal preferences or improve inference in some way. Improper priors have played an important part in many studies for reasons other than being convenient and commonly employed representations of ignorance. Some priors, such as the Jeffreys' prior, have information theoretic justifications and invariance properties, while others result in admissible or at least low (frequentist) risk estimators important for practical exercises such as forecasting or impulse response analysis. Being able to use some of these priors when calculating posterior model probabilities would allow us to retain these benefits while accounting for model uncertainty. However, since Bartlett (1957) it has generally been accepted that improper priors on all of the parameters result in ill-defined Bayes factors and posterior probabilities that prefer (with probability one) the smaller model regardless of the information in the data. This is commonly termed Bartlett's paradox and results because the ratio of prior normalising constants, $\mathfrak{c}_j/\mathfrak{c}_i$, is not well defined. For practice, Bartlett's paradox implies improper priors are used only for the common (to all models) parameters and proper priors must be specified for the remaining parameters when computing posterior model probabilities. A recent example of this principle is Fernández, Ley and Steel (2001) and further examples of authors comfortable with this approach are listed in Kass and Raftery (1995). The adoption of this principle has precluded the general use of improper priors in computing posterior probabilities.

In this paper we present a simple result which demonstrates that the class of priors that may be used to obtain posterior probabilities is wider than previously thought

and includes some improper priors. We do this by demonstrating that Bartlett's paradox does not hold for all improper priors - contrary to conventional wisdom. Decomposing the parameter vector into its norm and a unit vector, we provide a new representation of Bartlett's paradox in terms of the rate of divergence of the measure for the norm. We then use this representation in two further ways. First, we demonstrate that the improper Shrinkage prior results in well defined Bayes factors and, second, we use a nesting argument to demonstrate another prior that results in well defined Bayes factors and has properties similar to some priors already in use. The common feature of these priors is that their measure is a polynomial in the norm whose order is not a function of the dimension of the model.

We emphasise that it is *not* the primary aim of this paper to advocate the use of improper priors as another method of obtaining inference on model uncertainty that may be regarded as objective or as a reference approach. Having established the rather simple result that some improper priors do produce well defined Bayes factors, our second aim is to extinguish any hope that these might provide a trouble-free reference or objective prior for model comparison. We do this in two ways. Using the Jeffreys prior as an example, we discuss a limitation of the method used to prove the result. Next, with a discussion of the importance of the role of the prior measure in model comparison – which is lost when improper priors are used - we demonstrate a pathology that the above priors introduce into the Bayes factor.

Finally, we present a simple approach to regaining both a well defined Bayes factor and the features of the improper prior by controlling the *relative* size of the supports for parameters of different models. This approach places no real restriction upon inference as the supports are made arbitrarily large enough to ensure the posterior integrals have converged.

Much of the literature on BMA in econometrics has focused upon the Normal linear regression model with uncertainty in the choice of regressors (for a good introduction to this large body of literature, see Fernàndez, Ley and Steel 2001). Another contribution of this paper, therefore, is to extend the class of models and problems that may be considered with BMA. For much of the discussion we leave the form of alternative models largely unspecified except for their dimensions. We demonstrate an application of the priors to a relatively complex but economically useful set of models. This application gives some indication of the relative performance of the alternative priors and treatments of the prior measure.

The structure of the paper is as follows. In the following subsection we discuss approaches to obtaining model inference with improper priors as well as 'minimal information' or reference priors that have been presented in the literature. In Section 2 we outline the explanation for why the posterior distribution is well defined when a Uniform prior measure for the parameters with unbounded support is employed, while the Bayes factors are not. Section 3 provides an explanation for why some improper priors on common parameters only can be used to produce well defined Bayes factors and posterior probabilities. As mentioned, 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. Using the approach developed in this discussion, we demonstrate how some improper priors on all parameters result in well defined Bayes factors.
In Section 4 we discuss the Jeffreys prior to demonstrate both a limitation on the focus we take and show how the role of the prior measure for the parameter space is affected by the form of the priors discussed. Here we introduce an approach to using proper priors on supports of arbitrarily large diameter such that the Bayes factors are informed by the data and easily obtained, and link these to the use of particular improper priors. In Section 6 these priors are applied to a simple empirical example relating to the term structure of Australian interest rates. Section 7 contains some concluding comments and suggestions for further research.
Some notation for vector spaces and measures on these spaces with be useful for use in developing the discussion. The background theory is found in Muirhead (1982) (for further discussion 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 $\mathbb{O}(r)=\{C(r\times r):C'C=I_r\}$, that is $C\in\mathbb{O}(r)$. We distinguish this from the notation for the large order of magnitude of a sequence $\{c_i\}_{i=1}^{M}$ such that $d^{-n}c_i<K<\infty$, which we write as $O(d^n)$. The $n\times r$ $(n\geq r)$ 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 $v\in V_{1,n}$. When we refer to the diameter of a space A we refer to $d=diam(A)=\sup\{|x-y|:x,y\in A\}$ which may or may not be finite. Any $n\times 1$ vector $\theta\in A\subseteq R^n$ can be decomposed as $\theta=v\tau$ where τ is a positive scalar. We term a function of θ, $f(\theta)$, symmetric - as defined in Phillips (1994) - if $f(\theta)=f(v,\tau)=f(\tau)$. 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.

Theorem 1 *For $i=1,2$, let the $n\times 1$ vector θ_i have support $A_i\subseteq R^n$ and $f_i(\theta_i)$ be a symmetric function. If $c_i=\int_{A_i}f_i(\theta_i)\,d\theta_i$ diverges in d_i such that $c_i=O(d_i^n)$, then if $c_1=O(d^n)$ and $c_2=O(d^n)$, then $\frac{c_1}{c_2}=O(1)$.*

Proof. The proof follows from the basic properties of divergent sequences. If $x=O\left(m^h\right)$ and $y=O(m^g)$, then $z=xy=O\left(m^{h+g}\right)$. Let $x=c_1$ and $y=\frac{1}{c_2}$ and the result follows.
Important examples of such functions of central interest in this paper are

$$\alpha_n^d=\int_0^d\tau^{n-1}d\tau=\frac{d^n}{n}=O(d^n)$$

and

$$\beta_n^d=\int_0^d\nu^{n-1}d\nu=\frac{d^n}{n}=O(d^n).$$

We will use variants of the rather simple result (in this simple case, the Theorem 1 can be proved using l'Hopital's rule).

$$\lim_{d\to\infty} \frac{\alpha_n^d}{\beta_n^d} = \lim_{d\to\infty} \frac{\int_0^d \tau^{n-1} d\tau}{\int_0^d \nu^{n-1} d\nu} = \lim_{d\to\infty} \frac{nd^n}{nd^n} = 1. \tag{1}$$

Further we will use the result where for $q > 0$

$$\lim_{d\to\infty} \frac{\alpha_{n+q}^d}{\alpha_n^d} = \infty. \tag{2}$$

Despite the apparent simplicity of these results, their implications for model comparison with improper priors seem to have been overlooked.
Before we present the main result, the following Subsection gives a very brief overview of some of the literature on this topic and the variety of approaches that have been developed to deal with it. Due to the importance of this issue - as demonstrated by the calibre of authors that have attempted to address it in some way - this literature has become quite extensive and we do not pretend to give it a complete treatment. Rather, we highlight that nowhere has it previously been discussed that the issue (Bartlett's paradox) that generated this body of work is not a general as perceived.

1.1 Related literature

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 priors with the general aim of producing posterior model probabilities that contain no subjective prior information. An early approach to developing an approximation to the Bayes factors with minimal prior information is presented by Schwarz (1978) who uses an asymptotic argument to let the data dominate the prior as the sample size increases. For a fixed sample size in the linear model with Normal priors, Klein and Brown (1984) use limits of measures of information based upon those developed by Shannon (1948) to formalise the concept of 'minimising information'. Interestingly, for the particular model and prior they consider, they obtain the same expression as Schwarz to approximate the posterior odds ratio. These approaches assume proper priors, but use limiting arguments to allow the information in the sample to dominate that prior information.
A significant advance in asymptotic theory of Bayesian model selection by estimation of the marginal likelihood is made in Phillips and Ploberger (1996) and Phillips (1996). These papers also consider approximations to the marginal likelihood for a wide class of likelihoods and priors, again using asymptotic domination of the prior by the data, but they extend the class of models to those that include possibly nonstationary time series data, discrete and continuous data as well as multivariate models.
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 Fernández, 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 is to use proper priors so that we maintain a valid interpretation of the posterior. 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 Normal linear regression model which allow such comparison of results. They use improper priors on the common parameters - the intercept and the variance - and a Normal prior on the remaining coefficients based upon the g-prior of Zellner (1986). This approach is supported by the argument of Lindley (1997) - who used model comparison as one motivating example - that only proper priors should be employed to represent uncertainty.

Each of the methods discussed to this point have either removed the prior from the calculation of posterior probabilities or been limited in the class of prior or model or both. A notable alternative which requires neither the full likelihood nor a prior is the Bayesian Method of Moments proposed by Zellner (1994, 1997a, 1997b) and Zellner and Tobias (2001). As we have argued, some improper priors have attractive properties and do result in well defined Bayes factors and posterior probabilities. One approach with improper priors is given in Kleibergen (2004) using 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 linear regression model. A feature common to both Klein and Brown (1984) and Kleibergen (2004) is that the prior model probabilities are given limiting behaviour that offsets the divergent term in the Bayes factor (resulting in well defined Bayes factors). Kleibergen (2004) presents an approach that holds for a very general form for the prior but only where models nest, while the approach of Klein and Brown (1984) and the result we present are only relevant for specific forms of the prior. The result in this paper is more general in the sense that we make no assumptions about the forms of the models or their relationship to each other. The result does not require models to nest, nor does it place any restriction upon the specification of the prior probabilities for the models. As far as we are aware, this is a direction that has not been considered previously in the literature.

2 The posterior and Bartlett's paradox

In this section we provide an alternative representation of Bartlett's paradox. To do this, we begin with a discussion of the definition of the posterior with improper priors

as this explanation is well understood, generally accepted, and leads directly to an understanding of the paradox and of why some improper priors result in well defined Bayes factors. We also provide a justification for the common practice of using the same improper priors on common parameters (such as variances and intercepts) when computing posterior model probabilities and this provides an interpretation for our main result.

Let the n vector of parameters θ have support defined by $\theta \in \Theta \subseteq R^n$ with $\lambda(\Theta) = \infty$. We ignore parameters with compact supports with finite Lebesgue measure as they do not generally cause problems with the interpretation of the Bayes factor. Therefore when we refer to a model having a particular dimension, we mean by this the dimension of the space Θ of the model. Recall the prior density on θ is $\pi(\theta) = h(\theta)/\mathfrak{c}$ where $\mathfrak{c} = \int h(\theta)\, d\theta$ and the likelihood function is $L(y|\theta)$, the posterior density is defined as

$$\pi(\theta|y) = \frac{L(y|\theta)\,\pi(\theta)}{\int_\Theta L(y|\theta)\,\pi(\theta)\,d\theta} = \frac{L(y|\theta)\,h(\theta)/\mathfrak{c}}{\int_\Theta L(y|\theta)\,h(\theta)\,d\theta/\mathfrak{c}} = L(y|\theta)\,h(\theta)/p \tag{3}$$

where $p = \int_\Theta L(y|\theta)\,h(\theta)\,d\theta$. Even if we use an improper prior such as with $h(\theta) = 1$ and $\lambda(\Theta) = \infty$ so that $\mathfrak{c} = \infty$, the posterior is considered well defined (see for example Kass and Raftery 1995 or Fernández *et al.* 2001) so long as the integral p converges. We assume this is the case throughout the paper such that we only consider proper posteriors.

We restrict ourselves in the remainder of this section to the Uniform prior as used in Bartlett's original example as this is sufficient to demonstrate the issue and provides a useful base upon which we can build to investigate the properties of alternative prior measures.

Say we wish to investigate the properties of a vector of data y where we have two or more models. Denote model i by M_i and the n_i vector of parameters for this model as θ_i. The posterior probability of the model is given by $\Pr(M_i|y)$ and 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, $B_{ij} = p_i/p_j \times \mathfrak{c}_j/\mathfrak{c}_i$. The data inform the Bayes factor through the $p's$ and if the two models are considered *a priori* equally likely, the posterior odds ratio is equal to the Bayes factor. As our interest is in the influence of the prior on the Bayes factor, of real importance for our discussion is the ratio of the unnormalised prior measures for the parameter spaces for the two models, $\mathfrak{c}_j/\mathfrak{c}_i$. 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.

The ratio $\mathfrak{c}_j/\mathfrak{c}_i$ reflects our relative prior measure for Θ_j to that for Θ_i and plays an

important role in the Bayes factor by providing a relative weighting that accounts for the dimensions and diameters of the supports for the two models. This ratio incorporates a penalty for the relative dimensions as well as our uncertainty about the parameter values. Greater dimension or prior uncertainty about the model parameters will tend to increase $\mathfrak{c}$. For example, if we reflect greater prior uncertainty by a larger prior variance and give θ a multivariate Normal prior density with zero mean and covariance $\sigma^2 I_n$, then the prior measure for the space is $\mathfrak{c} = (2\pi)^{\frac{n}{2}} \sigma^n$. $\mathfrak{c}$ will therefore increase with dimension n and uncertainty σ. In many circumstances, restricting the diameter of the support might be regarded as reflecting a measure of certainty in place of σ. Another way of looking at σ is as a measure of the 'effective' diameter of the support. That is, beyond a certain distance measured in number of $\sigma's$, the values of θ contribute little to the prior mass. This is effectively a nondogmatic prior restriction on the support diameter. A general observation about the relationship between the normalising constant, $\mathfrak{c}$, and the dimension and measures of certainty, n and σ respectively, is that $\frac{\partial \mathfrak{c}}{\partial n} > 0$ and $\frac{\partial \mathfrak{c}}{\partial \sigma} > 0$. This relationship holds for a wide range of distributions commonly used for priors e.g., Normal, Wishart, Inverted Wishart. We will return to this role of the prior measure later in the paper.

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 the Bayes factor is also infinite and not well defined. In this case the penalty for uncertainty is absolute such that $\Pr(M_i|y) = 1$ and $\Pr(M_j|y) = 0$. But these posterior probabilities are not well defined in the sense that their values do not reflect any information in the data, only prior uncertainty. 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 only upon the relative dimensions of the two models. In the first and last cases in which the same degree of prior uncertainty is expressed, the posterior probabilities will assign probability one to the smallest model and zero to all other models considered such that the penalty for dimension is absolute. In each of these cases the data are unable to inform the posterior probabilities. The exception when $\mathfrak{c}_j/\mathfrak{c}_i = 1$ (see Poirier 1995 and Koop 2003) holds when the dimensions of the models match.

As these same results can be shown to occur with other improper priors, and regardless of whether one regards this as a paradox or a natural outcome in probability of using improper priors, there is clearly then a limitation to inference when employing improper priors. The conventional wisdom is that improper priors cannot be used for model comparison by posterior probabilities.

One generally accepted exception to the conventional wisdom is as follows. If we partition θ_k into (γ_k, γ) where γ are common to all models, we can show in the case where improper priors of the same form are used only on γ (of course the prior for θ_k is then improper. When we say that improper priors are only used on γ, we mean that the prior for γ_k conditional upon γ is proper), the Bayes factors will be well defined (see for example, Fernández *et al.*, 2001). In this case $\mathfrak{c}_k = \mathfrak{c}_{\gamma_k}\mathfrak{c}_\gamma$ where

$\mathfrak{c}_{\gamma_k} = \int h_k(\gamma_k|\gamma)(d\gamma_k) \leq M < \infty$ and $\mathfrak{c}_\gamma = \int g(\gamma)\, d\gamma = \infty$ thus $\mathfrak{c}_j/\mathfrak{c}_i = \mathfrak{c}_{\gamma_j}/\mathfrak{c}_{\gamma_i}$ since the $\mathfrak{c}_\gamma$ cancels. This result could be thought of as the basis of this paper as we reparameterise to isolate a common parameter, the norm of θ, upon which an improper prior is used. However, this in no way requires that the interpretation of the norms are the same, rather only that they have the same support, R^+. In fact, the supports need not be the same. Rather they need only be unbounded above some finite value for each model. This value need not be the same for any two models.

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 $\theta_i = v_i\tau_i$ where the $n_i \times 1$ vector v_i is a unit vector, $\nu_i'\nu_i = 1$, which defines the direction of θ_i and $\tau_i \geq 0$ defines the vector length. The vector v_i is an element of a Stiefel manifold V_{1,n_i}, $v_i \in V_{1,n_i}$. The compact space V_{1,n_i} has a measure $dv_1^{n_i}$ and volume

$$\varpi_{n_i} = \int_{V_{1,n_i}} dv_1^{n_i} = 2\pi^{n_i/2}/\Gamma(n_i/2) < \infty \tag{4}$$

(Muirhead, 1982). We can therefore decompose the differential term for θ_i into $d\theta_i = \tau_i^{n_i-1}(d\tau_i)\, dv_1^{n_i}$. The expression for the differential term leads to the following explanation for Bartlett's paradox. We can decompose the integral $\mathfrak{c}_i$ into a convergent (finite) part, ϖ_{n_i}, and the divergent part, α_{n_i}:

$$\mathfrak{c}_i = \int_{R^{n_i}} d\theta_i = \int_{R^+} \tau^{n_i-1}(d\tau) \int_{V_{1,n_i}} dv_1^{n_i} = \alpha_{n_i}\varpi_{n_i} \tag{5}$$

where

$$\alpha_{n_i} = \int_{R^+} \tau^{n_i-1}(d\tau) = \infty. \tag{6}$$

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 $\mathfrak{c}_j = \int_{R^{n_j}} d\theta_j = \alpha_{n_j}\varpi_{n_j}$. 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 $n_i = n_j = n$ and $h_i(\theta_i) = h_j(\theta_j) = 1$, then the Bayes factor $B_{ij} = m_i/m_j = p_i/p_j \times \mathfrak{c}_j/\mathfrak{c}_i$ where since $\mathfrak{c}_i = \mathfrak{c}_j = \alpha_n\varpi_n$, $B_{ij} = p_i/p_j$ is well defined since by (1) $\mathfrak{c}_j/\mathfrak{c}_i = 1$. The important point here is that we have taken the ratio of two polynomials (in the respective norms) of the same order such that they diverge at the same rate. 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.

Note that the integrals α_n and ϖ_n do not depend upon the chosen model, only its dimension, n. Further, provided the support of θ is unbounded in at least one direction, the term α_n is not affected by restrictions upon the support of θ. This is because such restrictions to $\Theta \subset R^n$ will restrict the support of v (not τ) and so restrict only the measure of this support, ϖ_n. For example, m positivity constraints (say for

variances) will reduce ϖ_n to $2^{-m}\varpi_n$. A possible and rather strange exception is if Θ_i is made up of a closed convex space around the origin and some other unbounded space such that, say, $\tau \in (0, u(v)] \times (l(v), \infty)$ for some $l > u$. However, it is the rate of divergence of the integral with respect to τ that results in Bartlett's paradox and this rate will not change. We can show this by replacing the lower bounds of the integrals for τ in (1) and (2) by positive finite numbers. The limits of the integrals and their ratios are unchanged.
When $n_j > n_i$, the integrals of τ (the term α_n) diverge at different rates and we have the case in (2) such that the ratio $\alpha_{n_j}/\alpha_{n_i} = \infty$. The term in B_{ij} due to the polar part will always be finite and known with value

$$\varpi_{n_j}/\varpi_{n_i} = \pi^{(n_j - n_i)/2} \frac{\Gamma(n_i/2)}{\Gamma(n_j/2)}. \tag{7}$$

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). It is conceivable then that by building upon the Uniform prior measure we may find other improper prior measures exist which result in a divergent part of the integral, the α_n, that diverges at the same rate for all models using this prior such that the ratio $\alpha_{n_j}/\alpha_{n_i}$ is finite (usually one) and B_{ij} is well defined. This is effectively using a common form of improper prior on τ. We present some examples in the following section.

3 Improper priors with well defined Bayes factors: exceptions to Bartlett's paradox

In this section we present the first result of the paper: the improper priors which result in well-defined Bayes factors exist. As has been discussed, many researchers accept that using improper priors on common parameters does not result in Bartlett's paradox. Here we show that in treating the norm of the parameter vector as a common parameter, certain improper priors on all parameters result in well defined Bayes factors.

3.1 The improper Shrinkage prior: normalising the differential term

The Shrinkage prior has been 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 does result in 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 $(\theta'\theta)^{1/2} = \tau$. The differential form of the prior is

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

and this form holds for all models. Importantly this prior results in a first order polynomial in τ for all models. The normalising constant for a model of dimension n is then

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

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, $\mathfrak{c}_j/\mathfrak{c}_i = \varpi_{n_j}/\varpi_{n_i}$ given in (7), is finite and known.

3.2 Nested prior

A number of methods developed for inference have nested models within a 'largest' model to produce sensible prior measures for the nested models. Kleibergen (2004) gives a careful specification of how to restrict from an encompassing model to an encompassed model, with examples, in such a way that the posterior odds are well defined even with improper priors. Using only proper priors, Fernández *et al.* (2001) point out that priors for nested models can be obtained from a prior on the full model so long as the priors (for the variance) for the nested models incorporate the term $(n - n_i)/2$ to account for the difference between the dimension of the largest model, n, and the nested model, n_i. We use a similar approach here with an improper Uniform prior on the largest model but also provide a formal justification in the Appendix I.
As 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, in fact, require the models to nest, nor do we restrict the augmenting parameter in the same way, however clearly nested models can be

accommodated. Despite the fact that the models need not nest, we call this a nested prior for two reasons. First, the form of the prior is developed by an argument that uses a nesting of the dimensions. Second, the most common comparisons in econometrics tend to be among models that nest within some encompassing model. Therefore, this provides an alternative to the approach developed by Kleibergen (2004) for nesting models.

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 $\theta_2 = \{\theta_0', \theta_1'\}$ where θ_1 is a n_1-dimensional vector. The measure for the prior $h(\theta) = h(\theta_2) = 1$ is given in (5) as $\mathfrak{c} = \alpha_n \varpi_n$. To obtain the measure for θ_0 in the model M_0 we give it the vector of parameters θ_2 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 resulting prior on θ_0 is $(\theta_0'\theta_0)^{n_1/2}(d\theta_0)$ (see Appendix I) with measure $\mathfrak{c}_0 = \alpha_n \varpi_{n_0}$. As shown in the Appendix, the ratio of the normalising constants becomes

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

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.

As the proof of the above result uses a 'conditioning upon a measure zero event' argument, it is necessary to comment upon an important paradox which arises in this case: the Borel-Kolmogorov paradox. Our comment is deliberately brief and restricted to stating why this paradox is not really an issue in the above case. The Borel-Kolmogorov paradox is encountered when different representations of the same measure zero event appear in different parameterisations. With the transformation from (θ_0, θ_1) to (θ_0, φ) where $\varphi = \varphi(\theta_0, \theta_1)$ the transformation of measures is $\nu(\theta_0, \theta_1) = \nu_{0|1}(\theta_0|\theta_1)\, \nu_1(\theta_1) = \varepsilon(\theta_0, \varphi) = \varepsilon_{0|\varphi}(\theta_0|\varphi)\, \varepsilon_\varphi(\varphi)$.

The Borel-Kolmogorov paradox implies that even if $\theta_1 = 0 \Longrightarrow \varphi = c$, it is not always true that $\nu_{0|1}(\theta_0|\theta_1 = 0) = \varepsilon_{0|\varphi}(\theta_0|\varphi = c)$. However, the case we give involves a vector θ_0 of model parameters and a vector θ_1 of artificial parameters. Any transformations that might sensibly be considered would be of θ_0, $\varphi_0 = \varphi_0(\theta_0)$, not $\varphi = \varphi(\theta_1)$. Thus we have $\nu_{0|1}(\theta_0|\theta_1 = 0) = \varepsilon_{\varphi_0|1}(\varphi_0|\theta_1 = 0)$ and the paradox does not arise. While it is not out of the question that some transformation could be imagined that involved both θ_0 and θ_1, it is difficult to imagine how such a transformation could be regarded as sensible. The vector θ_1 is purely artificial and does not enter into the model. Notwithstanding the comments above, the result presented does not depend upon the justification given.

Improper priors have measures that are divergent in the dimension of the support, d,

and it is this divergence that usually results in ill-defined Bayes factors. The priors we have presented above are, of course, just examples of a wide range of possible (improper) priors that could result in well defined Bayes factors. To produce well defined Bayes factors, the essential feature a prior need possess is a (possibly divergent in d) prior measure that is a polynomial in d of known degree that matches in the numerator and denominator of the Bayes factor.

4 Limitations and an alternative approach

In this section we discuss issues related to the analysis of improper priors using the above results including some important limitations of using these improper priors for model comparison and pathologies they introduce into the Bayes factor. These pathologies lead us to, first, recommend against using these priors in the current form and, second, to propose an alternative approach to employing improper priors.

4.1 The analysis of nonsymmetric priors: The Jeffreys prior for the Normal linear model

In the above discussion we have focussed upon the term in the prior measure associated with the norm τ with unbounded support, as this term resulted in the divergent component in the integral. However, it was possible to ignore the term involving the unit vector v only because the above priors are symmetric. Nonsymmetric priors present a limitation on this analysis as we must also consider the measure for v.
One important example is the Jeffreys prior for the multivariate Normal linear model $y = X\beta + \varepsilon$ in which y is a $T \times m$ random data matrix, X is the $T \times k$ matrix of regressors, β is a $k \times m$ matrix of unknown coefficients and $vec(\varepsilon) \sim N(0, \Sigma \otimes I_T)$. The symmetric covariance matrix $\Sigma = T'T$ is positive definite and T is the upper triangular Choleski decomposition of Σ with the $(i,j)^{th}$ nonzero element denoted as t_{ij}. We will denote the i^{th} diagonal element as t_{ii} and note $t_{ii} > 0$. Collect the $n = km + m(m+1)/2$ parameters into the $n \times 1$ vector $\theta = \left(vec(\beta)', vech(T)'\right)'$ with decomposition $\theta = \nu\tau$ with ordering for notational convenience such that $t_{ii} = v_{ii}\tau$.
We assume that the dimension of the system m is fixed and any zero restrictions of interest will be upon β or on the covariances in the off diagonal of Σ (if we consider, for example, certain exogeneity restrictions). This excludes the case where one or more variances are involved in linear restrictions (such as equalling zero). The following results are quite general as they will hold in all but this rather exceptional case.
The exact Jeffreys prior is the square root of the information matrix which in this case has the form (see Appendix II for the results in this section)

$$\begin{aligned} p(\beta,\Sigma)\, d(\beta,\Sigma) &\propto |\Sigma|^{-(k+m+1)/2}\, d(\beta,\Sigma) = \\ &= 2^m \Pi_{i=1}^m t_{ii}^{-(k+i)} d(\beta,T) = 2^m \Pi_{i=1}^m v_{ii}^{-(k+i)} dv_1^n \tau^{-1} d\tau. \end{aligned} \tag{8}$$

The prior measure for the parameter space will be $\mathfrak{c}_n = \int d\theta = 2^m \widetilde{\varpi}_k^n \alpha_0$ where $\widetilde{\varpi}_k^n = \int_{v_1^n} \Pi_{i=1}^m v_{ii}^{-(k+i)} dv_1^n$. Thus all models will have the term α_0 which will cancel in the Bayes factor, however $\widetilde{\varpi}_k^n$ is a divergent integral which results in ill-defined Bayes factors. The divergence results from the limits of the integrals in the regions where the v_{ii} approach zero and the rate of divergence is governed not only by k - the dimension of β and most frequently the object of interest - but also by the dimension n. This last point means if two models differ by one in the number of regressors, or even if two models do have the same number of regressors but a covariance (say exogeneity) restriction imposed, then integrals $\int_{v_1^n} \Pi_{i=1}^m v_{ii}^{-(k+i)} dv_1^n$ and $\int_{v_1^{n-1}} \Pi_{i=1}^m v_{ii}^{-(k+i)} dv_1^{n-1}$ diverge at different rates (The differing rates of divergence result from the dependence of the ν_{ii} upon the other ν_{ij} through the constraint $\nu'\nu = 1$. So keeping even k constant does not result in common rates of divergence if the covariances are restricted). Thus adaptations of priors that result in polynomials in the norm of matching order will not remove this divergence.

The effect of the divergence in $\widetilde{\varpi}_k^n$ could be removed and Bayes factors computed if we were to restrict the elements of the unit vector v for the variances, the v_{ii}, to have positive minimums $c_i > 0$. As the i^{th} variance can be expressed as $\sigma_i^2 = \Sigma_{j=1}^i t_{ji}^2 = \tau^2 \Sigma_{j=1}^i v_{ji}^2$ and the support of τ is unrestricted, this restriction on v_{ii} would not imply a restriction upon the marginal support of each element of θ, however, the supports would no longer be variation free. If we consider the case $m = 1$, for example, large values of β would mean a larger lower bound upon $\sigma^2 = v_{k+1}^2 \tau^2$ since $\tau^2 = \theta'\theta = \sigma^2 + \beta'\beta$. Of course, as the conditional posterior distribution for σ^2 in this model will tend to have little mass around zero for large values of β, this is not likely to be a serious restriction. The question of choice of c_i, however, remains.

We conducted a number of simulations to determine values of c_i that gave values of $\widetilde{\varpi}_k^n$ that might result in useful Bayes factors. Although more work needs to be done in this direction to gain a clearer picture of the implications of this restriction, we were able to get an early impression of the effect of varying c_i. Our conclusion is, however, that the penalty in the prior measure for being large remains very significant such that there will remain *too strong* a preference in the Bayes factors for small models which is overcome only if there is considerable support in the data for larger models.

Before we conclude this subsection, we mention the most commonly used form of the Jeffreys prior which is the approximation suggested by Jeffreys himself. This prior assumes independence of β and Σ and has the form

$$p(\beta, \Sigma)\, d(\beta, \Sigma) \propto |\Sigma|^{-(m+1)/2}\, d(\beta, \Sigma) = 2^m \Pi_{i=1}^m t_{ii}^{-i} d(\beta, T) =$$
$$= 2^m \Pi_{i=1}^m v_{ii}^{-i} dv_1^n \tau^{km-1} d\tau$$

In this case $\mathfrak{c}_n = \int d\theta = 2^m \widetilde{\varpi}_k \alpha_{km}$ where $\widetilde{\varpi}_k = \int_{v_1^n} \Pi_{i=1}^m v_{ii}^{-i} dv_1^n$ is still a divergent integral and depends upon n (and so k) so will not cancel in the Bayes factor. Further, the term α_{km} now enters which will result in the smallest model being selected.

This subsection demonstrates a clear limitation upon the result that prior measures

with matching orders of polynomials in the norm will not always produce computable Bayes factors. Careful consideration must be given to the how ν enters the prior.

4.2 The role of the prior measure

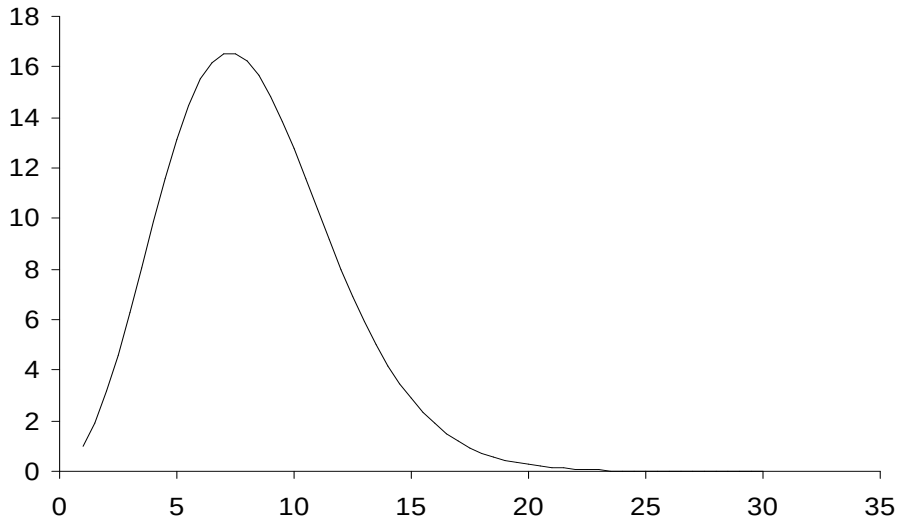
In this subsection we discuss the practical implications of using the improper priors that result in well defined Bayes factors discussed earlier. We present a pathology associated with these priors that suggests we should not in fact refer to the Bayes factors as 'well defined', but rather as 'able to be calculated'. The main problem is that an important function of the prior measure is lost with these improper priors. We conclude with a suggestion for proper priors that retains the attractive features of the improper priors but also result in well behaved Bayes factors.

As discussed in Section 2, with proper priors the ratio $\mathfrak{c}_j/\mathfrak{c}_i$ brings into the posterior analysis penalties for greater model dimension and greater prior parameter uncertainty. With the Shrinkage and Nested improper priors, the penalty for uncertainty is removed (effectively matched for each model). The ratio is then only a function of the dimensions of the models via the ratio $\mathfrak{c}_j/\mathfrak{c}_i = \varpi_{n_j}/\varpi_{n_i}$. Interestingly, this same ratio would result if we were to use a bounded spherical support centred at the origin of arbitrarily large diameter d such that all integrals $p_j = \int_{\Theta_j} L(\theta_j|y)\, h(\theta_j)\, d\theta_j$ have converged for all models. This same ratio would also result if we were to use Uniform proper priors over a spherical support centred at the origin and of arbitrarily large diameter d_i, but where we chose the diameters by the rule $d_i^{n_i}/n_i = d_j^{n_j}/n_j$ or $d_j = \left(\frac{n_j}{n_i} d_i^{n_i}\right)^{1/n_j}$. Note we need only choose the smallest d_i to be some arbitrarily large number such that all of the integrals p_j have converged. Thus we never need to actually assign a value to d_i, so long as we incorporate into the Bayes factor the correct value $\varpi_{n_j}/\varpi_{n_i}$. Of course these cases do not produce the same Bayes factor as the ratios p_i/p_j will differ, but they provide useful comparisons for discussion. For the Uniform prior, this choice of d_j ensures that as $d_i \to \infty$ the models with larger dimension have smaller diameter for the support.

This choice of a common limit on the norm (or a common rule for choosing d_j in the case of the Uniform prior) for all models is therefore innocuous in this case and holds as $d_i \to \infty$. Choosing d_j by such rules to remove the effect of the divergent part of the prior measure may seem like a useful simplification, however this process results in posterior odds with odd and undesirable properties.

It has become accepted that models of larger dimension should be penalised in the posterior via the prior measure. However, because of the behaviour of the ϖ_n over n, the penalty for dimension with the improper priors discussed is largely inverted as smaller models tend to be more heavily penalized. Figure (1) plots ϖ_n for $n = 1, ..., 30$, and shows the measure for $V_{1,n}$ is not monotonic in n, increasing up to around $n = 9$ and decreasing thereafter. The effect on the ratio $\mathfrak{c}_j/\mathfrak{c}_i = \varpi_{n_j}/\varpi_{n_i}$ is shown in Figure (2) which plots $\ln(\varpi_{gn}) - \ln(\varpi_n)$ for $n = 1, 2, 3, 4$ and 5 and $g = 1, ..., 20$. Recall that the larger the prior measure for a model, the more a model is penalized. Thus the more

Figure 1: Plot of ϖ_n, the measure for $V_{1,n}$, for $n = 1, ..., 30$.



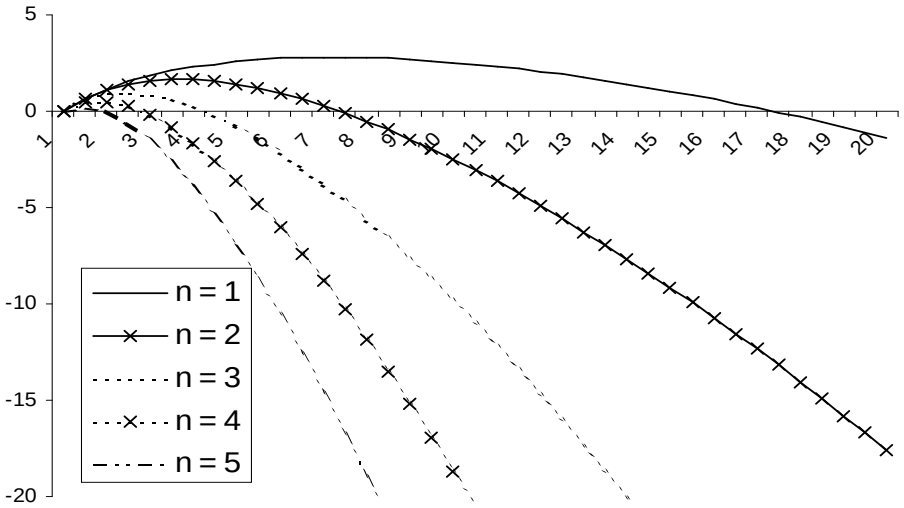
negative is $\ln(\varpi_{gn}) - \ln(\varpi_n)$ the greater is the penalty for the model of dimension n relative to the model of dimension gn. We see from Figure 2 then, that very small models (small n) are given less penalty than slightly larger models (small $g > 1$), but are heavily penalized relative to very large models (large g). As the dimension of the numerator (in the Bayes factor) model M_i increases, the penalty for being relatively small becomes very large very quickly.

This pathology is due to the non-monotonicity of ϖ_n in n. This effect is usually overwhelmed in improper priors by the integral with respect to the norm.

It would seem sensible, therefore, to use a different rule for selecting d_i. It is not recommended that the prior measures be completely ignored or dropped by assuming $\mathfrak{c}_j/\mathfrak{c}_i = 1$, however, as the role this ratio plays in the model selection or comparison is then unfulfilled. Ideally we would prefer a term that reintroduces a penalty for the dimension of the model, with a smooth increase in the measure as n increases, but results in a well defined term in the Bayes factor that does not give unmitigated support for the smallest (or largest) model.

As d_i increases, the normalising constant $\mathfrak{c}_i$ for an improper prior will usually diverge at a rate governed by d_i and n_i. However, as we will see, this is not always the case and different measures will diverge at different rates depending upon model dimension. Fortunately we may choose the relative diameters ($d = d_j/d_i$) by a rule such that the ratio $\mathfrak{c}_j/\mathfrak{c}_i$ is a function only of the relative dimensions n_j and n_i and provides a sensible penalty for model dimension. For example, for the Uniform measure on a

Figure 2: Plot of $\ln(\varpi_{gn}) - \ln(\varpi_n)$ for $n = 1, 2, 3, 4$ and 5 and $g = 1, ..., 20$. The value g is on the $x - axis$.



spherical support centered at the origin and of diameter d_i, then

$$\mathfrak{c}_i = \frac{\varpi_{n_i} d_i^{n_i}}{n_i}.$$

Say we choose d_i by the rule

$$\mathfrak{c}_i = \frac{\varpi_{n_i} d_i^{n_i}}{n_i} = \delta_0 \delta_1^{\frac{n_i}{2}} \propto \delta_1^{\frac{n_i}{2}}$$

such that for all d (with sufficiently large d_i) we obtain the Bayes factor $B_{ij} = p_i/p_j \delta_1^{(n_j - n_i)/2}$. For the Shrinkage prior and Nested prior, we can specify similar rules such as

$$\mathfrak{c}_i = \frac{\varpi_{n_i} d_i^2}{2} \propto \delta_1^{\frac{n_i}{2}}$$

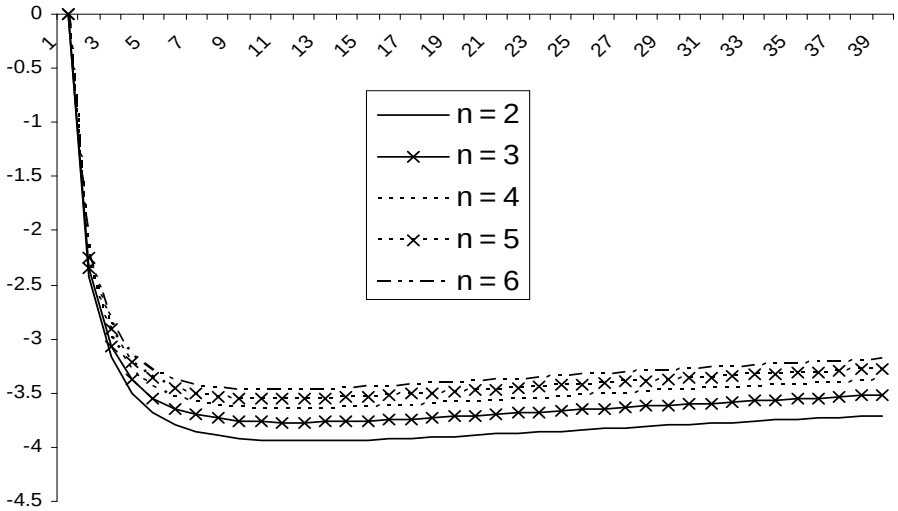
and

$$\mathfrak{c}_i = \frac{\varpi_{n_i} d_i^n}{n} \propto \delta_1^{\frac{n_i}{2}}$$

respectively. Thus for the Uniform prior, d increases as δ_0 increases and at a rate determined by n such that larger models have smaller diameter supports. As the size of the support reflects our certainty about the location of the parameter(s), we can regard larger supports as reflecting less certainty. Selecting a rule by which we determine the relative support sizes can therefore be viewed as a way of determining

the relative rate of increase in uncertainty. For the Shrinkage and Nested priors, these rules do not imply smaller supports for larger models. To explain this we need to refine our justification for the rules. Rather than controlling the support size directly, these rules control the relative uncertainty as measured by the weights in the Bayes factor given to the models of different dimensions, where this weight depends upon the rate of divergence of the chosen measure. To obtain sensible relative weights then, we sometimes need larger supports for larger models to allow them to accumulate sufficient volume.

Figure 3: This figure plots the log of the relative diameter d of the support for model of dimension n_i to the support for model of dimension $n_j = kn_i$, when using the rule for the Uniform prior. The n in the figure is n_i and $k = n_j/n_i$ is on the *x-axis*. Each line is for a different model dimension $n = n_i$. We have used $\delta_1 = T = 94$ to match with the application in Section 5. As we move from left to right, the dimension of the larger model increases. We have chosen $d_i = 1000$ for this case and increasing d_i moves the lines further down.



For sufficiently large supports (or large d_i), the integral p in (3) will have converged and can be estimated using standard approaches to approximate the marginal likelihood such as Chib (1995), Chib and Jeliazkov (2001) and Gelfand and Dey (1994). Rather than directly estimating the marginal likelihood, $m = p/c$, the product of the marginal likelihood and the prior measure is computer. For example, using Gelfand and Dey (1994) with known c and $q(\theta)$ such that $1 = \int q(\theta)\, d\theta$, we

Figure 4: This figure plots $\ln(d)$ when using the rule for the Shrinkage prior. See Figure 3 for information on how to interpret this figure.

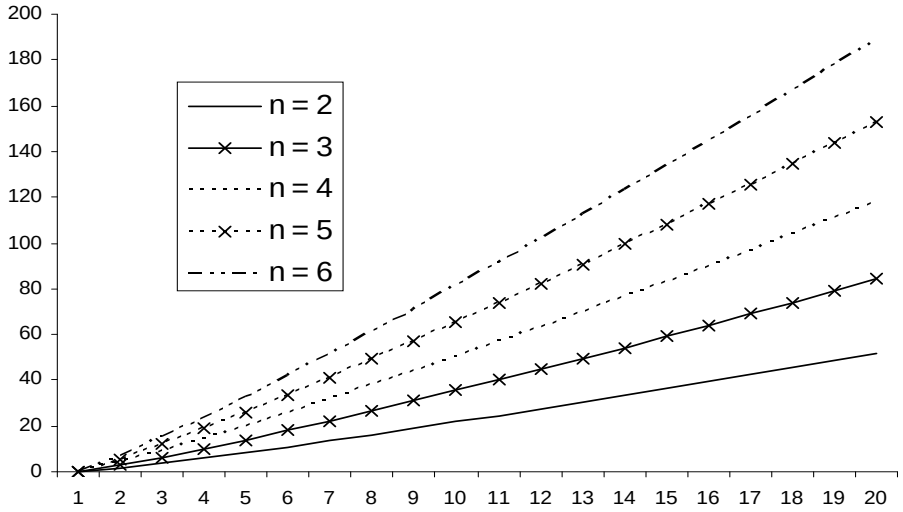
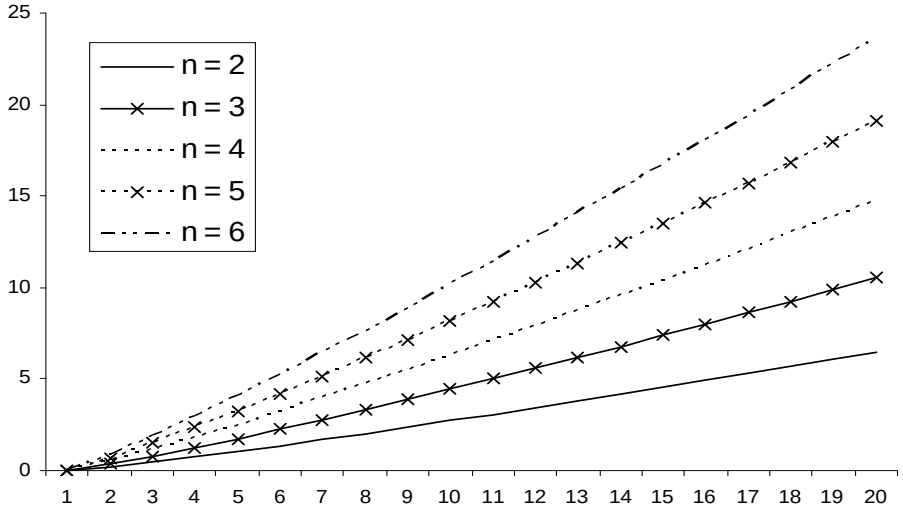


Figure 5: This figure plots $\ln(d)$ when using the rule for the Nested prior. See Figure 3 for information on how to interpret this figure. We have chosen as the maximum $n = 16$.



would compute

$$\frac{1}{m} = \int \frac{q(\theta)}{L(y|\theta)\,\pi(\theta)} p(\theta|y)\,d\theta.$$

With only the ratio c_i/c_j defined, we compute the product $p = mc$ as

$$\frac{1}{p} = \frac{1}{mc} = \int \frac{q(\theta)}{L(y|\theta)\,h(\theta)} p(\theta|y)\,d\theta.$$

Now we turn to the practical matter of assigning a value to δ_1. Two possible choices for δ_1 suggest themselves from the literature. The first is $\delta_1 = \pi$ as suggested in Kleibergen and Paap (2002, p. 238), and this will be equivalent to the choice of Chao and Phillips (1999) in computation of their posterior information criterion. Second, we could choose $\delta_1 = T$ the sample size, such that we obtain $B_{ij} = p_i/p_j T^{(n_j - n_i)/2}$. For the Uniform prior, this converges as $\delta \to \infty$ to the posterior odds ratio suggested by Klein and Brown (1984) and replaces the prior measure for the parameter space with the penalty used by Schwarz (1978) in his asymptotic approximation to the marginal likelihood. Thus this choice is equivalent to using the proper Uniform prior of arbitrarily large diameter where the relative diameters are chosen to match the unnormalised prior measures to the ratio of the BIC penalties.
If we have two models of dimensions n_i and $n_j = kn_i$, we can express the relative diameters of the supports, $d = d_j/d_i$, implied by the above rules as follows:

$$d = \left(\tfrac{T}{\pi}\right)^{(k-1)/2k} d_i^{(1-k)/k} \left(\frac{k\Gamma(kn_i/2)}{\Gamma(n_i/2)}\right)^{1/kn_i} \quad \text{for the Uniform prior;}$$

$$d = \left(\tfrac{T}{\pi}\right)^{(k-1)n_i/4} \left(\frac{\Gamma(kn_i/2)}{\Gamma(n_i/2)}\right)^{1/2} \quad \text{for the Shrinkage prior; and}$$

$$d = \left(\tfrac{T}{\pi}\right)^{(k-1)n_i/2n} \left(\frac{\Gamma(kn_i/2)}{\Gamma(n_i/2)}\right)^{1/n} \quad \text{for the Nested prior.}$$

From these expressions, we can show $\frac{\partial d_j}{\partial d_i} = \frac{d}{k} > 0$ for the Uniform prior, and $\frac{\partial d_j}{\partial d_i} = d > 0$ for the Shrinkage and Nested priors. To give a feel for the relative diameters of parameter spaces implied by the above method, we have plotted in Figures 3, 4 and 5 the log of the implied relative diameter d against the model dimension n_i for various models $k = n_j/n_i$. That is, these figures plot $\ln(d) = \ln(d_j) - \ln(d_i)$ where the index i denotes the smaller model. Figure 3 shows the log relative diameters $\ln(d)$ implied by the Uniform prior, Figure 4 shows $\ln(d)$ for the Shrinkage prior and Figure 5 shows $\ln(d)$ for the Nested prior. We have used $\delta_1 = T = 94$ to match with the application below.
For the Uniform prior, we see that the support diameter d_j of the larger model (with dimension $n_j = kn_i$) becomes very much smaller than the support diameter d_i for the smaller model as k increases up to around $k = 9$ or 10. After this point, the support diameter does not become smaller and actually increases. However, this increase is extremely slow. For example, for a model with diameter $d_i = 1000$ and

dimension $n_i = 6$, the diameter for a model 20 times larger ($n_j = 120$) is 3.37% of that for model i, but for a model 55 times larger ($n_j = 330$) this has only increased to 4.71%. The opposite effect takes place with the Shrinkage and Nested priors as the support diameters increase for larger models. This is necessary to offset the effect demonstrated in Figure 2.
We have come back to recommending the use of proper priors. However, the above suggestion allows us to maintain the features of certain improper priors that bring particular benefits to inference such as reducing frequentist risk. As this recommendation requires only a decision on the relative dimensions of the supports, and not on the actual dimension of any one support, all we essentially require is a method of computing or estimating p_i as if the support were unbounded. We conclude this section by making the point that the above method works only for divergent measures on unbounded supports and so will not be practical (or at all necessary) for proper priors: i.e., with convergent measures on unbounded supports.

5 Application

In this section we investigate evidence on the rational expectations theory for the term structure of interest rates (Campbell and Shiller, 1987) in which we expect that 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 may have non-zero means, we would not expect there to be trends in either the levels or the cointegrating relations. We use a vector error correction model (VECM) which has several other features about which we are uncertain. We use a $p = 4$ dimensional time series vector, $y_t = (y_{1t}, \dots y_{pt})$ for $t = 1, \dots, T$. The data for this example is 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) and Strachan and van Dijk (2003).
With a maximum of 3 lags and differencing, we have an effective sample size of $T = 90$ observations. The VECM of the $1 \times p$ vector time series process y_t, conditioning on the l observations $t = -l + 1, \dots, 0$, is $\Delta y_t = y_{t-1}\beta\alpha + d_t\mu + \Sigma_{i=1}^{l}\Delta y_{t-i}\Gamma_i + \varepsilon_t$. The matrices β and α' are $p \times r$ and assumed to have rank r. We will define $d_t\mu$ shortly. Collect the above parameters, except β, into

$$b = \left(vec(\alpha)', vec(\mu)', vec(\Gamma_1)', \dots, vec(\Gamma_l)'\right)'.$$

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 and r is thus obvious and in the following paragraphs we explain the parameterisation of models with different d and o.
We consider a range of deterministic processes such that Δy_t may have a nonzero mean or trend (implying a drift in y_t) and $y_t\beta$ may have a nonzero mean or trend. For specification of the restrictions that induce these behaviours we refer to Johansen (1995 Section 5.7). 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. For the interest rate data, we would most likely expect $d = 4$ or $d = 5$.

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
$E(\Delta y_t)$	$\mu_1 + \delta_1 t$	μ_1	μ_1	0	0
$E(y_t\beta)$	$\mu_0 + \delta_0 t$	$\mu_0 + \delta_0 t$	μ_0	μ_0	0

The aim of cointegration analysis is essentially to determine the dimension (r) and the direction of the cointegrating space, $\rho = sp(\beta)$. We therefore compare three models for the spaces of interest. When no restriction is placed upon the space and ρ is free to vary over all of the Grassman manifold 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 $o = 1$.
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 $l \in [0, 1, 2]$, and the indicator for overidentification of cointegrating vectors $o \in [1, 2, 3]$. This gives a total of 225 models. Taking account of observationally equivalent or *a priori* impossible models, we need only compute the marginal likelihoods for some 135 models.
The prior for β is uniform on $V_{r,p}$ but we adjust the volume to imply a uniform prior on the support of the cointegrating space (see Strachan and Inder 2004 for details). The same prior for the covariance matrix, the invariant partial Jeffreys prior for Σ, $p(\Sigma) \propto |\Sigma|^{-(p+1)/2}$, is employed for all models. For i^{th} model the prior for the n_i-dimensional vector b is $p(b) \propto (b'b)^{K_i/2}$ where $K_i = n^* - n_i$ where $n^* = \max(n_h)$ for the prior using Nested model which augments the differential, and $K_i = -(n_i - 2)$ for the Shrinkage prior. The marginal likelihoods are estimated by the MCMC approach of Strachan and van Dijk (2004) which uses such approaches as those discussed in Gelfand and Dey (1994).

Table 1: Estimated Posterior Model Probabilities (only values $\geq 1\%$ shown)

				$(b'b)^{-(n-2)/2}$	$(b'b)^{(n^*-n)/2}$	$(b'b)^{-(n-2)/2}$	$(b'b)^{(n^*-n)/2}$	*Prior*
d	l	r	o	ϖ_n	ϖ_n	$T^{-n/2}$	$T^{-n/2}$	*Penalty*
4	1	1	1	0.07	0.06	0.075	0.06	
5	1	1	1	0.28	0.94	0.287	0.94	
5	1	1	2	0.03	-	0.035	-	
5	1	1	3	0.59	-	0.597	-	

Table 1 shows the results from Bayesian estimation from the Shrinkage prior ($(b'b)^{-(n-2)/2}$) and the Nested prior ($(b'b)^{(n^*-n)/2}$) and where we have used the unbounded support to obtain the Bayes factor (ϖ_n) and the bounded supports with the rules for diameter as described in the previous section to obtain the Bayes factor ($T^{-n/2}$). Overall the results prefer models with low order or no deterministic processes, no lags of differences and three common stochastic trends. The evidence on the overidentifying restrictions is less clear with the Nested prior preferring the least restricted model while the Shrinkage prior shows a slight posterior preference (the posterior odds for $o = 1$ to $o = 3$ for the shrinkage prior is 2 which is not generally regarded as strong evidence. See for example, Kass and Raftery (1997), Poirier (1995) or Jeffreys (1961).) for the most restricted, although with considerable support (around 35%) upon the least restricted model.

This result gives clear evidence for this data set against the main feature of the Efficient Market Hypothesis that the interest rates share a single common stochastic trend, although there is some support that the spreads are stationary within each market. This model provides a reasonable description of the deterministic and short-run dynamic structure.

Although we have not used a particularly large sample, 90 observations seem to have been sufficient to dominate the effect of the form of the prior and the penalty for dimension in what is a reasonably complex model set. Interestingly, the form of the correction to the Bayes factor, either ϖ_n or $T^{-n/2}$, does not seem to have had much effect upon the results. Further, although we would expect that such different priors as the Shrinkage and the Nested priors to produce different results - with the Shrinkage prior preferring smaller models - again this did not produce great differences except for the restrictions upon the cointegrating space. Although we used a common prior in all cases for the cointegrating space, ρ, and we assumed prior independence of b and ρ, it is not surprising that the prior on b will affect inference in the posterior upon ρ since the two are not independent in the posterior which has a different form under each prior.

6 Conclusions

Due to Bartlett's paradox, Bayesians have not believed it possible to employ improper priors when obtaining posterior probabilities for models. This is unfortunate as some improper priors have attractive features which the Bayesian may like to employ in, say, BMA. Using a relatively simple and well-understood decomposition of the differential term for a vector of parameters, we have demonstrated that certain improper priors do result in well defined Bayes factors. One important class is the Shrinkage prior which has been shown 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 to those with other attractive properties. This is a potential area for further investigation.

While we present two classes of priors that do produce well defined Bayes factors, we show that these resulting Bayes factors are not well behaved. The problem is the relative prior measures which bias posterior inference in favor of larger models. From a discussion on the role of the prior measure in model selection or model weighting, we present an method of using the same form as the improper prior distributions but on a compact space - a sphere of given diameter centered at the origin - such that the prior is now proper. The approach essentially sets rules for determining the relative sizes of support diameters for models of different dimensions in such a way that the role of the prior measure in the Bayes factor is restored. Importantly, however, the actual size of the support diameters are unspecified and can be arbitrarily large so that they play no further role in the computation of the Bayes factor. We can therefore select the ratio of prior measure to be something that reflects what we judge to be reasonable penalties for increased dimension.

7 Acknowledgements

The authors are grateful to Arnold Zellner for very useful discussion on the topic of the paper.

References

[1] Bartlett, M. S. (1957) A comment on D.V.Lindley's statistical paradox. *Biometrika* **44**, 533-534.

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

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

[4] Bernardo, J.M. (1979) Expected information as expected utility. *The Annals of Statistics* **7**, 686-690.

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

[6] Chao, J. C. & P. C. B. Phillips (1999) Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics* **91**, 227-271.

[7] Chib, S. and I. Jeliazkov (2001) "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270-281.

[8] Fernández, C., E. Ley & M. F. J. Steel (2001) Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**, 381-427.

[9] Gelfand, A.E., & D. K. Dey (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B* **56**, 501–504.

[10] Jeffreys, H. (1961) *Theory of Probability 3rd ed.* Oxford: Clarendon Press.

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

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

[13] Kass, R. E. & A. E. Raftery (1995) Bayes Factors. *Journal of the American Statistical Association* **90**, 773-795.

[14] Kleibergen, F. (2004) Invariant Bayesian inference in regression models that is robust against the Jeffreys-Lindley's paradox. *Journal of Econometrics* **123**, 227-258.

[15] Kleibergen, F. & R. Paap (2002) Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration. *Journal of Econometrics* **111**, 223-249.

[16] Klein, R. W. & S. J. Brown (1984) Model selection when there is minimal prior information. *Econometrica* **52**, 1291-1312.

[17] Koop, G (2003) *Bayesian Econometrics.* John Wiley and Sons Ltd, England.

[18] Leonard, T. & Hsu, J. S. J. (2001) *Bayesian Methods.* Cambridge: Cambridge University Press.

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

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

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

[22] Magnus, J. R. & H. Neudecker (1988) *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons, New York.

[23] Min, C. & Zellner, A., (1993) Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics* **56**, 89-118..

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

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

[26] Ni, S. X. & D. Sun (2003) Noninformative priors and frequentist risks of Bayesian estimators of vector-autoregressive models. *Journal of Econometrics* **115**, 159-197.

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

[28] Phillips, P. C. B. (1996) Econometric model determination. *Econometrica* **64**, 763–812.

[29] Phillips, P. C. B. & W, Ploberger (1996) An asymptotic theory of Bayesian inference for time series. *Econometrica* **64**, 381-412.

[30] Poirier, D. (1995) *Intermediate Statistics and Econometrics: A Comparative Approach.* Cambridge: The MIT Press.

[31] Raftery, A.E., D. Madigan & J. A. Hoeting (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179-191.

[32] Shannon, C. E. (1948) A mathematical theory of communication. *The Bell System Technical Journal* **27**, 378–423.

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

[34] Sclove, S. L. (1968) Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association* **63**, 596-606.

[35] Sclove, S.L. (1971) Improved estimation of parameters in multivariate regression. *Sankhya, Series A* **33**, 61-66.

[36] Spiegelhalter, D. J. & 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.

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

[38] Strachan, R. W. & H. K. van Dijk (2003) Bayesian model selection with an uninformative prior. *Oxford Bulletin of Economics and Statistics* **65**, 863-876.

[39] Strachan, R. W. & B. Inder (2004) Bayesian analysis of the error correction model. *Journal of Econometrics* **123**, 307-325.

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

[41] 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.

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

[43] Stein, C. (1962) Confidence sets for the mean of a multivariate Normal distribution. *Journal of the Royal Statistical Society, Series B* **24**, 265-296.

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

[45] Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P.K., Zellner, A. (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*. North-Holland, Amsterdam, 233-243.

[46] Zellner, A. (1994) Bayesian method of moments (BMOM) analysis of mean and regression models. In: Lee, J., Johnson, W., Zellner, A. (Eds.), *Prediction and Modelling Honoring Seymour Geisser*. Springer, New York, 61–74.

[47] Zellner, A. (1997a) Bayesian analysis in econometrics and statistics: The Zellner view and papers. In: Perlman, M., Blaugh, M. (Eds.), *Economists of the 20th Century Series*, Edward Elgar, Cheltenham, UK.

[48] Zellner, A. (1997b) The Bayesian method of moments (BMOM): Theory and applications. *Advances in Econometrics*. In: Fomby, T., Hill, R. (Eds.), *Applying Maximum Entropy to Econometric Problems*, Vol. 12. Jai Press, Greenwich, CT, 85–105.

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

[50] Zellner, A. & J. Tobias (2001) Further results on the Bayesian method of moments analysis of the multiple regression model. *International Economic Review* 42, February.

[51] Zellner, A. & W. A. Vandaele (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.

A Appendix I

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

$$V = \left[\begin{array}{cc} v & V_\perp \end{array} \right] \text{ where } v = \left[\begin{array}{c} v_0 \\ v_1 \end{array} \right] \text{ and } V_\perp = \left[\begin{array}{cc} V_{00,\perp} & V_{01,\perp} \\ V_{10,\perp} & V_{11,\perp} \end{array} \right] \tag{9}$$

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 $(d\theta) = V'(d\theta) = V'v(d\tau) + V'(dv)\tau$ or

$$(d\theta) = \left[\begin{array}{c} v'v \\ V_\perp' v \end{array} \right] (d\tau) + \left[\begin{array}{c} v'(dv) \\ V_\perp'(dv) \end{array} \right] \tau = \left[\begin{array}{c} (d\tau) \\ V_\perp'(dv)\,\tau \end{array} \right]$$

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 partitioning $\widetilde{V}_\perp$ the same as $V_\perp$ in (9), we have at $v_1 = 0$,

$$\widetilde{V}_\perp' v = \left[\begin{array}{c} \widetilde{V}_{00,\perp}' v_0 + \widetilde{V}_{01,\perp}' v_1 \\ \widetilde{V}_{10,\perp}' v_0 + \widetilde{V}_{11,\perp}' v_1 \end{array} \right] = \left[\begin{array}{c} \widetilde{V}_{00,\perp}' v_0 \\ \widetilde{V}_{10,\perp}' v_0 \end{array} \right] = 0.$$

This implies that at $v_1 = 0$, 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 $V_{11,\perp} = I_{n-n_0}$. That is, the space spanned by $\widetilde{V}_\perp$ will lie in the $n_1 = n - n_0$ 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 $n_0 = 2$. $v = (v_{11}, v_{21}, v_{31})'$ is a vector in a three dimensional space and each element of the vector relates to one coordinate. 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 restricted to the two dimensional plane defined by the first two coordinate axis. The matrix $\widetilde{V}_\perp$ now always lies in the plane passing through the third coordinate axis defined by the matrix $V_\perp = \begin{bmatrix} v'_{12} & v'_{22} & 0 \\ 0 & 0 & 1 \end{bmatrix}'$.
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

$$\begin{aligned}
(d\theta) &= V'(d\theta) = V'v(d\tau) + V'(dv)\tau = \\
&= \begin{bmatrix} v'_0 v_0 + v'_1 v_1 \\ V'_{00,\perp} v_0 + V'_{01,\perp} v_1 \\ V'_{10,\perp} v_0 + V'_{11,\perp} v_1 \end{bmatrix} (d\tau) + \begin{bmatrix} v'(dv) \\ V_{00,\perp}(dv_0) + V'_{01,\perp}(dv_1) \\ V_{10,\perp}(dv_0) + V'_{11,\perp}(dv_1) \end{bmatrix} \tau = \\
&= \begin{bmatrix} (d\tau) \\ V_{00,\perp}(dv_0)\tau \\ (dv_1)\tau \end{bmatrix} \text{ at } v_1 = 0 \text{ where } V_\perp = \begin{bmatrix} V_{00,\perp} & 0 \\ 0 & I_{n_1} \end{bmatrix}
\end{aligned}$$

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

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

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 $(d\theta) = \tau^{n-1}(d\tau)(dv_1^n)$ and $(d\theta_0) = \tau^{n_0-1}(d\tau)(dv_1^{n_0})$. Since $n = n_0 + n_1$ and $\theta_0'\theta_0 = \tau^2$, then clearly we have the same result if in the prior for M_0 we replace $(d\theta_0)$ by $(\theta_0'\theta_0)^{n_1/2}(d\theta_0) = = \tau^{n_1}\tau^{n_0-1}(d\tau)(dv_1^{n_0}) = \tau^{n-1}(d\tau)(dv_1^{n_0})$.

B Appendix II

Theorem: The exact Jeffreys prior for the multivariate Normal linear regression model has the form (see Appendix II)

$$\begin{aligned} & p(\beta,\Sigma)\, d(\beta,\Sigma) \propto |\Sigma|^{-(k+m+1)/2}\, d(\beta,\Sigma) = \\ & = 2^m \Pi_{i=1}^m t_{ii}^{-(k+i)} d(\beta,T) = \\ & = 2^m \Pi_{i=1}^m v_{ii}^{-(k+i)} dv_1^n \tau^{-1} d\tau. \end{aligned}$$

Proof: The multivariate Normal linear model has the form $y = X\beta + \varepsilon$ in which y is a $T \times m$ random data matrix, X is the $T \times k$ matrix of regressors, β is a $k \times m$ matrix of unknown coefficients and $vec(\varepsilon) \sim N(0, \Sigma \otimes I_T)$. The information matrix for $\widetilde{\theta} = \left(vec(\beta)', vech(\Sigma)'\right)'$ has the form

$$\Upsilon = \begin{bmatrix} \Sigma^{-1} \otimes X'X & 0 \\ 0 & \frac{T}{2} D_m' \left(\Sigma^{-1} \otimes \Sigma^{-1}\right) D_m \end{bmatrix}$$

(Magnus and Neudecker, 1988, p. 321). The determinant of this matrix is then

$$|\Upsilon| = \left|\Sigma^{-1} \otimes X'X\right| \left|\frac{T}{2} D_m' \left(\Sigma^{-1} \otimes \Sigma^{-1}\right) D_m\right| = |X'X|^m |\Sigma|^{-k} T^{\frac{m(m+1)}{2}} |\Sigma|^{-(m+1)}$$

in which we have used the result $\left|D_m \left(\Sigma^{-1} \otimes \Sigma^{-1}\right) D_m\right| = \left|D_m^+ (\Sigma \otimes \Sigma) D_m^{+\prime}\right|^{-1} = = 2^{\frac{m(m-1)}{2}} |\Sigma|^{-(m+1)}$ (Magnus and Neudecker 1988, p. 50).
As the square root of the determinant of the information matrix, the Jeffreys prior will therefore be proportional to $|\Sigma|^{-(k+m+1)/2} d(\beta,\Sigma)$. Next, from Muirhead (1982, p. 62) we have the transformation of the measure from Σ to T as $(d\Sigma) = 2^m \Pi_{i=1}^m t_{ii}^{m+1-i}(dT)$ and so

$$\begin{aligned} |T|^{-(k+m+1)} 2^m \Pi_{i=1}^m t_{ii}^{m+1-i}(dT)(d\beta) &= 2^m \Pi_{i=1}^m t_{ii}^{-(k+m+1)} \Pi_{i=1}^m t_{ii}^{m+1-i}(dT)(d\beta) = \\ &= 2^m \Pi_{i=1}^m t_{ii}^{-(k+i)}(dT)(d\beta). \end{aligned}$$

The transformation $\theta = \left(vec(\beta)', vech(T)'\right)' = v\tau$ implies $(dT)(d\beta) = d\theta = = dv_1^n \tau^{n-1} d\tau$ where recall $n = km + \frac{m(m+1)}{2}$. Therefore we can write the Jeffreys

prior for (v,τ) for this model as proportional to

$$\Pi_{i=1}^{m} v_{ii}^{-(k+i)} \tau^{-\left(km+\frac{m(m+1)}{2}\right)} dv_1^n \tau^{km+\frac{m(m+1)}{2}-1} d\tau = \Pi_{i=1}^{m} v_{ii}^{-(k+i)} dv_1^n \tau^{-1} d\tau.$$

Beginning with the approximation of the Jeffreys prior as $|\Sigma|^{-(m+1)/2} d(\beta,\Sigma)$ and transforming from Σ to T, this becomes

$$\begin{aligned} |T|^{-(m+1)} 2^m \Pi_{i=1}^{m} t_{ii}^{m+1-i} (dT)(d\beta) &= 2^m \Pi_{i=1}^{m} t_{ii}^{-(m+1)} \Pi_{i=1}^{m} t_{ii}^{m+1-i} (dT)(d\beta) = \\ &= 2^m \Pi_{i=1}^{m} t_{ii}^{-i} (dT)(d\beta). \end{aligned}$$

The transformation from θ to $v\tau$ gives us the Jeffreys prior for (v,τ) for this model as proportional to

$$\Pi_{i=1}^{m} v_{ii}^{-i} \tau^{-\frac{m(m+1)}{2}} dv_1^n \tau^{km+\frac{m(m+1)}{2}-1} d\tau = \Pi_{i=1}^{m} v_{ii}^{-i} dv_1^n \tau^{km-1} d\tau.$$

The Nested prior

$$\begin{aligned} &p(\beta_0,\Sigma)\, d(\beta_0,\Sigma) = p(\beta,\Sigma)|_{\beta_1=0} d(\beta,\Sigma)|_{\beta_1=0} \propto \\ &\propto |\Sigma|^{-(k+m+1)/2} (b_0' b_0)^{k_1 m/2} d(\beta,\Sigma)|_{\beta_1=0} \end{aligned}$$

can be decomposed as

$$\begin{aligned} p(\beta_0,\Sigma)\, d(\beta_0,\Sigma) &\propto 2^m \Pi_{i=1}^{m} t_{ii}^{-(k+i)} (dT) (b_0' b_0)^{k_1 m/2} (d\beta_0) \propto \\ &\propto \Pi_{i=1}^{m} v_{ii}^{-(k+i)} \tau^{-\left(km+\frac{m(m+1)}{2}\right)} dv_1^n \tau^{km+\frac{m(m+1)}{2}-1} d\tau \\ &= \Pi_{i=1}^{m} v_{ii}^{-(k+i)} dv_1^n \tau^{-1} d\tau \end{aligned}$$

which has the same form in ν and in τ such that the rates of divergence of the divergent components of the integral will match.
Using the form of the Shrinkage prior we have the decomposition

$$\begin{aligned} p(\beta_0,\Sigma)\, d(\beta_0,\Sigma) &\propto |\Sigma|^{-(k+m+1)/2} (b_0' b_0)^{-k_0 m/2} d(\beta_0,\Sigma) \propto \\ &\propto \Pi_{i=1}^{m} t_{ii}^{-(k+i)} (dT) (b_0' b_0)^{-k_0 m/2} (d\beta_0) \propto \\ &\propto \Pi_{i=1}^{m} v_{ii}^{-(k+i)} \tau^{-\left(km+\frac{m(m+1)}{2}\right)} dv_1^n \tau^{-k_0 m} \tau^{k_0 m+\frac{m(m+1)}{2}-1} d\tau \propto \\ &\propto \Pi_{i=1}^{m} v_{ii}^{-(k+i)} dv_1^n \tau^{-km-1} d\tau \end{aligned}$$

which again has the same form in ν and in τ such that the rates of divergence of the divergent components of the integral will match.