University of Leicester

DEPARTMENT OF ECONOMICS

IMPROPER PRIORS WITH WELL DEFINED

BAYES FACTORS

Rodney W. Strachan,
University of Leicester, UK

Herman K. van Dijk,
Erasmus University Rotterdam, The Netherlands

Improper priors with well defined Bayes Factors.

Rodney W. Strachan[1] and Herman K. van Dijk[2]

[1]Department of Economics, University of Leicester,

Leicester, L1 7RH, U.K. email: rws7@le.ac.uk

[2]Econometric Institute, Erasmus University Rotterdam,

Rotterdam, The Netherlands. email: hkvandijk@few.eur.nl

ABSTRACT

While some improper priors have attractive properties, it is generally claimed that Bartlett's paradox implies that using improper priors for the parameters in alternative models results in Bayes factors that are not well defined, thus preventing model comparison in this case. In this paper we demonstrate, using well understood principles underlying what is already common practice, that this latter result is not generally true and so expand the class of priors that may be used for computing posterior odds to two classes of improper priors: the shrinkage prior; and a prior based upon a nesting argument. Using a new representation of the issue of undefined Bayes factors, we develop classes of improper priors from which well defined Bayes factors result. However, as the use of such priors is not free of problems, we include discussion on the issues with using such priors for model comparison.

Key Words: Improper prior; Bayes factor; marginal likelihood; shrinkage prior; measure.

JEL Codes: C11; C52; C15; C32.

1 Introduction.

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 a that model, ignoring 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 important 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. 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

2

(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. 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 Fernándes, 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.

Our aim is to 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

3

demonstrate that the improper shrinkage prior results in well defined Bayes factors and, second, we develop a prior that results in well defined Bayes factors and has properties similar to some priors already in use. Using the commonly employed Jeffreys prior as an example, we discuss a limitation of the method used to prove the result.

We emphasise that it is *not* the primary aim of this paper to produce another method of obtaining inference on model uncertainty that may be regarded as objective or as a reference approach. In fact we provide in the discussion a caveat on the use of these improper priors relating to an important role of the prior measure for the parameters in model comparison that is lost when improper priors are used. We give a simple suggestion how to regain some of this benefit of proper priors.

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 Fernàndes, 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.

4

The structure of the paper is as follows. In Section 2 we outline the explanation for why the posterior distribution is well defined when a Uniform prior measure for the parameters with unbounded support is employed, while the Bayes factors are not. We also explain why some improper priors on common parameters can be employed in estimating posterior probabilities of the models and this may be regarded as the principle underlying the result in this paper. 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. In Section 3 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.

The main result is presented in Section 4 where the improper priors are developed. In Section 5 we provide discussion using the Jeffreys prior to demonstrate 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 $O(r) = \{C(r \times r) : C'C = I_r\}$, that is $C \in O(r)$. The $n \times r$ semi-orthogonal matrix V is an element of the Stiefel manifold denoted by $V_{r,n} = \{V(n \times r) : V'V = I_r\}$, that is $V \in V_{r,n}$. If $r = 1$, then V is a vector which we will denote by lower case such as v and $v \in V_{1,n}$. When we refer to the diameter of a space A we refer to $d = diam(A) = \sup\{|x - y| : x, y \in A\}$ which will be finite only if A is compact. 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.

An entity of central interest in this paper is $\alpha_n^d = \int_0^d \tau^{n-1} d\tau = \frac{d^n}{n}$ with limit $\alpha_n = \lim_{d \to \infty} \frac{d^n}{n} = \infty$ but we also use variants of the rather simple result

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

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 seems to have been overlooked.

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. If the prior density on θ is $\pi(\theta) = h(\theta)/\mathfrak{c}$ where $\mathfrak{c} = \int h(\theta) d\theta$ is the unnormalised prior measure for the parameter space, and the likelihood function is $L(\theta)$, the posterior density is defined as

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

where $p = \int_\Theta L(\theta) h(\theta) d\theta$. Even if we use an improper prior such as with $h(\theta) = 1$ and $\lambda(\Theta) = \infty$ such that $\mathfrak{c} = \infty$, the posterior is considered well defined (see for example Kass and Raftery 1995 or Fernándes *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.

Consider the investigation of the properties of a vector of data y where we have two or more models and 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 given by $P(M_i|y)$ is a useful measure of the evidence in y for M_i. 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 we only consider proper posteriors (such that the ratio p_i/p_j will be well defined) and our interest is in Bartlett's paradox which is concerned with 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

8

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 weighting the support for the two models. This ratio incorporates a penalty for the relative dimensions as well as our uncertainty about the parameter values. Either of these features of the model will tend to increase $\mathfrak{c}$. For example, if we reflect greater prior uncertainty by a larger prior variance[1] and give θ a prior density with the form of a multivariate normal 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. A general observation, then, of relationship between the normalising constant, $\mathfrak{c}$, and the dimension, n, and measures of certainty, σ, 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.

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 ∞ so that the Bayes factor is also infinite and not well defined. In this case the penalty for uncertainty is absolute and $\Pr(M_i|y) = 1$ 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

---

[1]In many circumstances, restricting the diameter of the support might be regarded as reflecting a measure of certainty in place of σ.

9

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 poster 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 and 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,[2] the Bayes factors will be well defined as assumed in, for example, Fernándes *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

---

[2] 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.

$\mathfrak{c}_j/\mathfrak{c}_i = \mathfrak{c}_{\gamma_j}/\mathfrak{c}_{\gamma_i}$ 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 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^+.

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 $\tau_i > 0$ 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

$$\varpi_{n_i} = \int_{V_{1,n_i}} dv_1^{n_i} = 2\pi^{n_i/2}/\Gamma(n_i/2) < \infty \tag{3}$$

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

$$\mathfrak{c}_i = \int_{R^{n_i}} d\theta_i = \int_{R^+} \tau^{n_i-1}(d\tau) \int_{V_{1,n_i}} dv_1^{n_i} = \alpha_{n_i}\varpi_{n_i} \tag{4}$$

11

where

$$\alpha_{n_i} = \int_{R^+} \tau^{n_i - 1} (d\tau) = \infty. \tag{5}$$

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) $\mathfrak{c}_j/\mathfrak{c}_i = 1$. 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 and, provided the support of θ is unbounded in one direction, then the term α_n is generally not affected by restrictions upon the support. This is because restrictions to $\Theta \subset R^n$ will usually 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, this will not

change the important feature of the integral over τ which is the rate of divergence. This can be seen 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

$$\varpi_{n_j}/\varpi_{n_i} = \pi^{(n_j - n_i)/2}\frac{\Gamma\left(n_i/2\right)}{\Gamma\left(n_j/2\right)}. \tag{6}$$

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 an common form of improper prior on τ. We present some examples in Section 4. Before we present this result, the following section gives a very brief overview of the literature on this topic and the variety of approaches that have been developed to deal with it. This literature is quite extensive and we do not pretend to give it a complete treatment. We mention this literature to demonstrate the importance this topic has been given and the calibre of authors that have attempted to address it in some way. However, we emphasis again that we

13

do not aim to contribute to this wealth of approaches with another method. Rather, we present the result that the issue (Bartlett's paradox) that generated this body of work is not a general as perceived.

3 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 to 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 c_j/c_i may be replaced with estimates based upon some minimal amount of information from the sample. Examples of such approaches are Spiegelhalter and Smith (1982), O'Hagan (1995), and Berger and Pericchi (1996). This approach has an intuitive appeal and has been supported by asymptotic arguments. However, as discussed in Fernándes, Ley and Steel (2001), the use of the data to attribute a value to c_j/c_i involves an invalid conditioning such that the posterior cannot be interpreted as the conditional distribution given the data.

An alternative approach that has been proposed which maintains a valid interpretation of the posterior is to use proper priors. The rationale here is to compare Bayes factors for models with the same amount of prior information. To this end, Fernández *et al.* (2001) propose reference priors for the 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. 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). While Kleibergen (2004) presents an approach that holds for a very general form for the prior, the approach of Klein and Brown (1984) and the result we present (we do not present an approach) are only relevant for specific forms of the prior. However, this paper's result 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.

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

In this section we present the main result of the paper: the improper priors which result in well-defined Bayes factors. 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 parameters, certain improper priors result in well defined Bayes factors.

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

17

$\theta = v\tau$ such that $(\theta'\theta)^{1/2} = \tau$. Thus 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. The normalising constant for model M_i 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 will be $\varpi_{n_j}/\varpi_{n_i}$ which is given in (6) and is finite and known.

Augmenting the differential term.

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

18

$(n - n_i)/2$ to account for the difference between the dimension of the largest model, n, and the nested model, n_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 require the models to nest, nor do we restrict the augmenting parameter in the same way, however clearly nested models can be accommodated. 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 (4) 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

19

infinite Lebesgue measure. The resulting prior on θ_0 is $(\theta_0'\theta_0)^{n_1/2} (d\theta_0)$ (see Appendix I). 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 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 $\varphi = (\theta_0, \theta_1)$ with transformation of measures is

$$\nu(\theta_0, \theta_1) = \nu_{0|1}(\theta_0|\theta_1) \nu_1(\theta_1) = \varepsilon(\theta, \varphi) = \varepsilon_{1|\varphi}(\theta_1|\varphi) \varepsilon_\varphi(\varphi).$$

The paradox implies that even if $\theta_1 = 0 \implies \varphi = c$, it is not always true that $\nu_{0|1}(\theta_0|\theta_1 = 0) = \varepsilon_{1|\varphi}(\theta_1|\varphi = c)$. 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, $\varphi = \varphi(\theta_0)$, not $\varphi = \varphi(\theta_1)$. Thus we have $\nu_{0|1}(\theta_0|\theta_1 = 0) = \varepsilon_{\varphi|1}(\varphi|\theta_1 = 0)$ 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. The discussion on this point in the paper was given to provide some intuition for the result.

5   Discussion.

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.

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 parameter 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. When considering non-symmetric priors it is necessary to consider the terms involving v also.

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

21

of regressors, β is a $k \times m$ matrix of unknown coefficients and $vec\left(\varepsilon\right) \sim N\left(0, \Sigma \otimes I_T\right)$. 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} and so has i^{th} diagonal element $t_{ii} = v_{ii}\tau > 0$. Collect the $n = km + m\left(m+1\right)/2$ parameters into the $n \times 1$ vector $\theta = \left(vec\left(\beta\right)', vech\left(T\right)'\right)'$.

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) in which the following results are not valid. 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)

$$
\begin{aligned}
p\left(\beta, \Sigma\right) d\left(\beta, \Sigma\right) \quad &\propto \quad \left|\Sigma\right|^{-(k+m+1)/2} d\left(\beta, \Sigma\right) \\
&= \quad 2^m \Pi_{i=1}^m t_{ii}^{-(k+i)} d\left(\beta, T\right) = 2^m \Pi_{i=1}^m v_{ii}^{-(k+i)} dv_1^n \tau^{-1} d\tau.
\end{aligned}
\tag{7}
$$

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 resulting in ill-defined Bayes factors. The divergence results from the limits of the integrals in the regions where the v_{ii} approach zero. Here 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 even if two models 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.

If we consider using the embedding approach in the previous section, this can lead to a specification for the Jeffreys prior that results in well defined Bayes factors. Specify the Jeffreys prior for the augmented model with parameters (β, Σ) of dimension n as in (7). Partition β as $\beta = [\beta_0', \beta_1']'$ and we are interested in the nested model implied by setting the mk_1 elements of β_1 to zero. Next we consider the point where a subset of mk_1 elements of $\beta = [\beta_0', \beta_1']'$ are set to zero . If $vec(\beta_1)$ is a $k_1 \times 1$ vector, then at $\beta_1 = v_1 \tau = 0$, the form of the prior for (β_0, Σ) becomes

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

Alternatively, using the shrinkage argument and let $k_0 = k - k_1$, the prior

$$p(\beta_0, \Sigma) d(\beta_0, \Sigma) \propto |\Sigma|^{-(k+m+1)/2} (\beta_0'\beta_0)^{-k_0 m/2} d(\beta_0, \Sigma)$$

also results in well defined Bayes factors if k is set to some common value (such as the maximum) for all models. Notice that in neither of these cases an adjustment for the volume of the Steifel manifolds of differing dimensions relevant, since the integral over this space is also divergent.

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 ith 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 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 a lot of support in the data for larger models.

In concluding 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

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

24

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 but common to all models and so will cancel in the Bayes factor. However, the term α_{km} now enters which will result in the smallest model being selected.

The role of the prior measure.

We limit the aim of this paper to presenting the result that the Bayes factors are well defined for the priors considered, and not presenting a new model selection strategy because an important function of the prior measure is lost with these improper priors. 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 augmented 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 spherical support centred at the origin of arbitrarily large diameter d such that all integrals $p_j = \int_{\Theta_j} L(\theta_j) d\theta_j$ 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, so long as we adjust the Bayes factor by the correct

value $\varpi_{n_j}/\varpi_{n_i}$.

These cases is will not produce the same Bayes factor, however, as the ratios p_i/p_j will differ, but they provide useful comparisons for discussion. For the Uniform prior, this choice of d_i ensures that 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 in the case of the Uniform prior) for all models is therefore innocuous in this case and holds as $d \to \infty$. Choosing d by such rules to remove the effect of the prior measure may seem like a useful simplification, however this process results in posterior odds with odd and undesirable properties.

Because of the behaviour of the ϖ_n over n, the penalty for dimension with these priors 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 so that the more 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 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

26

the dimension of the numerator (in the Bayes factor) model M_i increases, the penalty for being small becomes very large very quickly.

It would therefore seem sensible to use a different rule for selecting d_i. It is not recommended that the prior measures be completely ignored or dropped by assuming $\frac{c_j}{c_i} = 1$, 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. Although detailed discussion of strategies to adjust for this loss of penalty is beyond the scope of this paper, we mention one that immediately suggests itself. That is to set d by the rule $c_i = \frac{d_i^{n_i}}{\varpi_{n_i} n_i} = \delta T^{\frac{n_i}{2}}$ such that for all d we obtain the Bayes factor $B_{ij} = p_i/p_j T^{(n_j - n_i)/2}$ where T is the sample size and so common to all models, but d increases as δ increases and at a rate determined by n such that larger models have smaller diameter supports. For the Uniform prior, this converges as $\delta \to \infty$ to the posterior odds ratio suggested by Klein and Brown (1984) and so 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 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.

6 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

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

28

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 $d = 5$.

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

The aim of cointegration analysis is essentially to determine the dimension (r) and the direction of the cointegrating space, $\rho = sp\left(\beta\right)$. 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 226 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 augmentation of 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).

d	l	r	o	$(b'b)^{-(n-2)/2}$ ϖ_n	$(b'b)^{(n^*-n)/2}$ ϖ_n	$(b'b)^{-(n-2)/2}$ $T^{-n/2}$	$(b'b)^{(n^*-n)/2}$ $T^{-n/2}$	*Prior* *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: Estimated Posterior Model Probabilities (only values of 1% shown)

Table 1 shows the results from Bayesian estimation from the shrinkage prior $((b'b)^{-(n-2)/2})$ and the augmenting prior $((b'b)^{(n^*-n)/2})$ and where we have used the exact form of the Bayes factor (ϖ_n) and the adjustment to account for model dimension $(T^{-n/2})$. 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 augmenting prior preferring the least restricted model while the shrinkage prior shows a slight posterior preference[3]

---

[3] The posterior odds for $o = 1$ to $o = 3$ for the shrinkage prior is 2 which is not generally regarded

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 that the spreads are stationary within each market has some support. 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 the exact Bayes factor or with the adjustment by $T^{-n/2}$ 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 augmenting 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.

as strong evidence. See for example, Kass and Raftery (1997), Poirier (1995) or Jeffreys (1961).

7 Conclusion.

Due to Bartlett's paradox, Bayesians have not employed 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.

8 References.

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

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

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

Bernardo , J.M. (1979) 'Expected information as expected utility' *The Annals of*

Statistics **7**, 686-690.

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

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

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

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

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

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

Kass, R. E. and A. E. Raftery (1995) 'Bayes Factors' *Journal of the American Statistical Association* **90**, 773-795.

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

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

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

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

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

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

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

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

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

Min, C. and 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..

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

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

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

O'Hagan, A. (1995) 'Fractional Bayes Factors for Model Comparison' *Journal of the*

Royal Statistical Society, Series B **57**, 99-138.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Zellner, A. (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.

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

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

9 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

$$V = \begin{bmatrix} v & V_\perp \end{bmatrix} \text{ where } v = \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} \text{ and } V_\perp = \begin{bmatrix} V_{00,\perp} & V_{01,\perp} \\ V_{10,\perp} & V_{11,\perp} \end{bmatrix} \quad (8)$$

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

38

or

$$(d\theta) = \begin{bmatrix} v'v \\ V_\perp'v \end{bmatrix}(d\tau) + \begin{bmatrix} v'(dv) \\ V_\perp'(dv) \end{bmatrix}\tau = \begin{bmatrix} (d\tau) \\ V_\perp'(dv)\tau \end{bmatrix}$$

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 (8), we have at $v_1 = 0$,

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

This implies that at $v_1 = 0$, 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

$$
\begin{aligned}
(d\theta) &= V'(d\theta) = V'v(d\tau) + V'(dv)\tau \\[2mm]
&= \begin{bmatrix} v'_0 v_0 + v'_1 v_1 \\ V'_{00,\perp} v_0 + V'_{01,\perp} v_1 \\ V'_{10,\perp} v_0 + V'_{11,\perp} v_1 \end{bmatrix} (d\tau) + \begin{bmatrix} v'(dv) \\ V_{00,\perp}(dv_0) + V'_{01,\perp}(dv_1) \\ V_{10,\perp}(dv_0) + V'_{11,\perp}(dv_1) \end{bmatrix} \tau \\[2mm]
&= \begin{bmatrix} (d\tau) \\ V_{00,\perp}(dv_0)\tau \\ (dv_1)\tau \end{bmatrix} \quad \text{at } v_1 = 0 \text{ where } V_{\perp} = \begin{bmatrix} V_{00,\perp} & 0 \\ 0 & I_{n_1} \end{bmatrix}
\end{aligned}
$$

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 $(\theta_0'\theta_0)^{n_1/2}(d\theta_0) = \tau^{n_1}\tau^{n_0-1}(d\tau)(dv_1^{n_0}) = \tau^{n-1}(d\tau)(dv_1^{n_0})$.

10   Appendix II

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

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

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 $k \times m$

41

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 $\left|D_m \left(\Sigma^{-1} \otimes \Sigma^{-1}\right) D_m\right| = \left|D_m^+ \left(\Sigma \otimes \Sigma\right) D_m^{+\prime}\right|^{-1} = 2^{\frac{m(m-1)}{2}} |\Sigma|^{-(m+1)}$ (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

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

42

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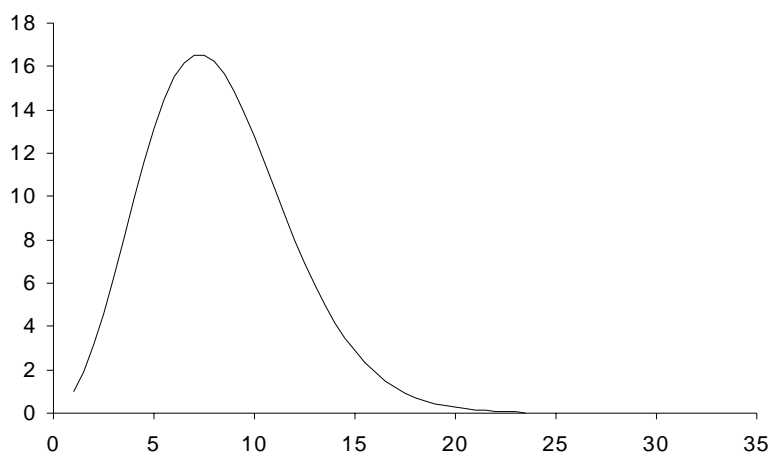
Figures

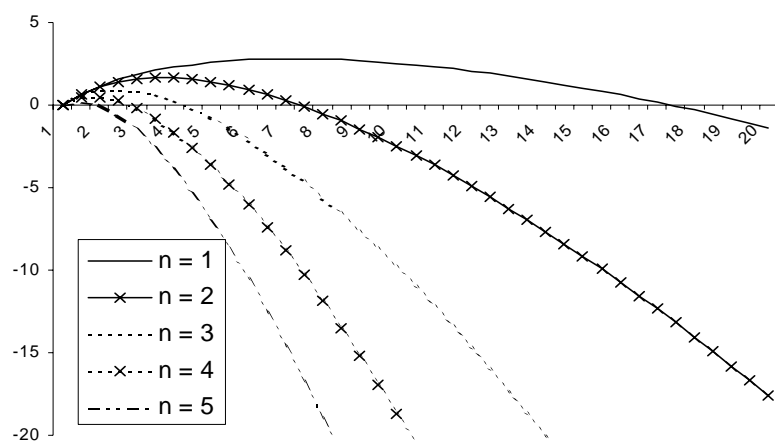Figure 1: Plot of ϖ_n, the measure for $V_{1,n}$, for $n = 1, ..., 30$.

Figure 2: Plot of $\ln\left(\varpi_{gn}\right) - \ln\left(\varpi_n\right)$ for $n = 1, 2, 3, 4$ and 5 and $g = 1, ..., 20$.