TI 2011-006/4
Tinbergen Institute Discussion Paper

Divergent Priors and Well Behaved Bayes Factors

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

[1] *Australian National University, Australia;*
[2] *Econometric Institute, Erasmus University Rotterdam, Tinbergen Institute.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

More DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Divergent priors and well behaved Bayes Factors.

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

[1]Research School of Economics, Australian National University, Australia

The Rimini Center for Economic Analysis

Centre for Applied Macroeconomic Analysis

email: rodney.strachan@anu.edu.au

[2]Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands.

email: hkvandijk@few.eur.nl

ABSTRACT

Divergent priors are improper when defined on unbounded supports. Bartlett's paradox has been taken to imply that using improper priors results in ill-defined Bayes factors, preventing model comparison by posterior probabilities. However many improper priors have attractive properties that econometricians may wish to access and at the same time conduct model comparison. We present a method of computing well defined Bayes factors with divergent priors by setting rules on the rate of diffusion of prior certainty. The method is exact; no approximations are used. As a further result, we demonstrate that exceptions to Bartlett's paradox exist. That is, we show it is possible to construct improper priors that result in well defined Bayes factors. One important improper prior, the Shrinkage prior due to Stein (1956), is one such

example. This example highlights pathologies with the resulting Bayes factors in such cases, and a simple solution is presented to this problem. A simple Monte Carlo experiment demonstrates the applicability of the approach developed in this paper.

Key Words: Improper prior; Bayes factor; marginal likelihood; Shrinkage prior; measure. **JEL Codes:** C11; C52; C15; C32.

1 Introduction

This paper addresses the issue of how to compute Bayes factors when priors diverge. In Bayesian analysis, Bayes factors or posterior probabilities play an important role for model comparison, model selection and model averaging. Certain improper priors are also important in Bayesian analysis for various reasons. For example some priors serve as representations of ignorance, or are used because of information theoretic justifications or invariance properties, or because they result in admissible or at least low (frequentist) risk estimators, or because they are used simply as a base case for determining the role of informative (usually proper) priors in an analysis. Since Bartlett (1957), however, it has been generally accepted that if improper priors are used the posterior probabilities are not well defined. Specifically, using improper priors results in posterior probabilities that prefer (with probability one) the smaller model regardless of the information in the data.[1] In this paper we consider improper priors as the limit of a sequence of priors that diverge in the diameter of the support. We have three aims. First, we present a method of obtaining well defined and well behaved Bayes factors with divergent priors by controlling the rate of diffusion of certainty. This may be thought of as using priors that are proper on compact supports of unspecified diameters, but these priors diverge in the diameter of the support such

[1] Bartlett does not only consider improper priors. He also considers divergent priors and the arbitrariness these introduce into the posterior probabilities.

that at the limit they become improper. Second, we establish classes of priors that are exceptions to Bartlett's paradox. Finally, we identify pathologies in the Bayes factors that result from using these improper priors and demonstrate how to use our first result to avoid these pathologies.

Before proceeding, it is important to distinguish between Lindley's paradox and Bartlett's paradox as the two are commonly confused. Lindley (1957) demonstrates how a given value of a test statistic contains different amounts of evidence on a hypothesis depending upon the sample size. Lindley (1957) does not consider improper priors. Bartlett (1957), in a comment on Lindley (1957), shows the effect of a diverging prior measure on the posterior odds and in particular the effect of an improper prior. In this paper we consider divergent priors and do not consider the effect of the sample size.

The structure of the paper is as follows. In Section 2 we define and generalize Bartlett's paradox. This discussion includes a formal justification for a number of standard practices such as using improper priors on common parameters and then computing Bayes factors and posterior probabilities. In Section 3 we introduce an approach to obtaining well defined exact Bayes factors from divergent priors.

Section 4 shows exceptions to Bartlett's paradox exist, with the specific example of the shrinkage prior of Stein (1956, 1960 and 1962). We demonstrate how some improper priors could result in well defined Bayes factors. However, we also show, in

Section 5, that the use of these priors implies possibly undesirable behaviour of the Bayes factor. Further, Section 5 contains general discussion on the Jeffreys (1961) prior which demonstrates an alternative way priors can diverge to result in ill-defined Bayes factors and shows how to adapt our suggested approach to account for this case. Section 6 presents a small Monte Carlo experiment to investigate the performance of three divergent priors in model selection. Section 7 contains some concluding comments and suggestions for further research.

2 Bartlett's paradox

In this section we restate Bartlett's paradox to provide a more general formulation than the original presentation. To do this, we begin with a 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 the reason 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 justification provides an interpretation for our main result. An important role of this section is to set up the techniques we use in the remainder of the paper.

Let $\lambda(A)$ denote the Lebesgue of the collection of spaces A, with $\lambda(A) = \infty$ implying that A has infinite Lebesgue measure. Next, let the n vector of parameters

θ have support defined by $\theta \in \Theta \subseteq R^n$ with $\lambda(\Theta) = \infty$. When we refer to a model having a particular dimension n, we mean by this the dimension of the space Θ of the model. Denote the prior distribution as $\pi(\theta) = h(\theta)/c$ where $c = \int h(\theta)\, d\theta$ is the unnormalized prior measure for the parameter space and $h(\theta)$ is a kernel for the density. If θ has unbounded support and $\pi(\theta)$ is improper then we regard c as not well defined and this is sometimes represented by the statement $c = \infty$. The likelihood function is $L(\theta|y)$ and the posterior density is defined as

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

where $p = \int_\Theta L(\theta|y)\,h(\theta)\,d\theta$. Even if we use an improper prior such as with $h(\theta) = 1$ and $\lambda(\Theta) = \infty$ so that $c = \infty$, 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. That the posterior is well defined is explained by assuming $c/c = \infty/\infty = 1$. It is only in specific circumstances that we can regard $\infty/\infty = 1$, and this is one case. However, we later discuss other cases in which limits of ratios of divergent measures (the $c's$) have finite limits.

Next we state a proposition for priors on bounded supports. This proposition will be directly useful later in the paper, but for now it provides a framework to permit us to consider unbounded supports. Some notation for vector spaces and measures on these spaces with be useful for use in developing the discussion. When we refer

to the diameter of a space A we refer to $d = diam\,(A) = \sup\left\{\|x-y\|^{1/2} : x, y \in A\right\}$. Any $n \times 1$ vector $\theta \in A \subseteq R^n$ can be decomposed as $\theta = v\tau$ where v is a unit vector such that $v'v = 1$ and τ is a positive scalar; $\tau > 0$. For the following proposition, we define a function of θ, $f(\theta)$, symmetric as defined in Phillips (1994). That is, for the decomposition $\theta = v\tau$, $f(\theta) = f(v, \tau) = f(\tau)$.

Assumption 1: For $i = 1, 2$, let the $n_i \times 1$ vector θ_i have support $A_i \subseteq R^n$ and let $f_i(\theta_i) > 0$ be a symmetric function. Assume the sequence $c_i = \int_{A_i} f_i(\theta_i)\, d\theta_i$ diverges in $d_i = diam\,(A_i)$. Given n_i and n_j, for every finite value of d_i there exists a finite value for d_j such that $\frac{c_1}{c_2} < \infty$.

The assumption restricts us to prior measures, c_i, that diverge in the diameter of the support, d_i. Since d_1 and d_2 go to the same limit and we are in the following case only interested in the limit, we can regard the sequences c_1 and c_2 as both diverging in $d = d_1 = d_2$. Treating the measure c_i for unbounded supports as the limit of a sequence in d we can state the following theorem.

Theorem 1 *For $i = 1, 2$, let the $n_i \times 1$ vector θ_i have support $A_i \subseteq R^{n_i}$ and $f_i(\theta_i)$ be a symmetric function. If the sequence $c_i = \int_{A_i} f_i(\theta_i)\, d\theta_i$ diverges in $d = d_i = diam\,(A_i)$ such that $c_i = O\left(d^{n_i}\right)$, then $\frac{c_1}{c_2} = O(1)$ if $n_1 = n_2$, $\frac{c_1}{c_2} = O\left(d^{n_1 - n_2}\right)$ if $n_1 > n_2$, and $\frac{c_1}{c_2} = o\left(d^{\delta + n_1 - n_2}\right)$ for some $\delta > 0$ if $n_1 < n_2$.*

Proof. The proof follows from the basic properties of divergent sequences. If $x = O\left(m^h\right)$ and $y = O\left(m^g\right)$, then $z = xy = O\left(m^{h+g}\right)$. Let $x = c_1$ and $y = \frac{1}{c_2}$ and

the result follows. ■

Theorem 1 shows how we can regard c/c as finite, but does not directly explain why we might legitimately assign the value $c/c = 1$. If, however, $c = O(d^n)$ then we can treat this as a polynomial in d and by *l'Hopital's* rule we obtain $c/c = 1$.

Assumption 2: For all models the support $A \subseteq R^n$ with diameter d and $p_A = \int_A L(\theta|y)\, h(\theta)\, d\theta$, is such that for any other support $\underline{A} \supset A$ with diameter $\underline{d} > d$ and $p_{\underline{A}} = \int_{\underline{A}} L(\theta|y)\, h(\theta)\, d\theta$, then $p_{\underline{A}} - p_A < \varepsilon$ for any $\varepsilon > 0$.

Assumption 2 essentially says that the diameters of the supports are all large enough such that increasing them results in no distinguishable increase in the posterior normalizing constant. This is a generalization of the same assumption made by Bartlett (1957) in the uniform prior single parameter case.

The above results will prove useful when we consider Bartlett's paradox. 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. To consider other types of priors, then we should replace the expressions such as $c = O(d^n)$ with $c = O(g(d))$ where $g(d)$ is monotonically increasing in d. We can also use the above results to consider priors that diverge on compact subsets of the real space. We show this when we consider the Jeffreys prior for the general linear regression model in Section 5.

Denote the prior distribution as $\pi(\theta) = h(\theta)/c$ where $c = \int h(\theta)\, d\theta$ is the unnormalized prior measure for the parameter space and $h(\theta)$ is a kernel for the density. If θ has unbounded support and $\pi(\theta)$ is improper then we regard c as not well defined.

Using the above results we give a more general statment of Bartlett's paradox. Say we wish to investigate the properties of a vector of data y where we have two or more potential 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 $m_i = p_i/c_i$ is the marginal density of y under model i. Therefore,

$$B_{ij} = \frac{p_i}{p_j} \times \frac{c_j}{c_i}.$$

The data inform the Bayes factor through the p_i and p_j and if the two models are considered a priori equally likely, the posterior odds ratio is equal to the Bayes factor. Our interest is in the influence of the prior on the Bayes factor through the ratio c_j/c_i. If a proper prior is used for each model such that $c_i < \infty$ and $c_j < \infty$ are well defined - and possibly known or able to be estimated - the Bayes factor is well defined as the ratio c_j/c_i is also defined.

If we use an improper prior with unbounded support for M_j and a proper prior for M_i, then $c_j = O\left(d_j^{n_j}\right)$ while $c_i = O\left(1\right)$ such that the ratio c_j/c_i diverges and the Bayes factor is infinite and not well defined. These posterior probabilities are not well defined in the sense that their values do not reflect any information in the data, only prior uncertainty. In this case the penalty for uncertainty is absolute such that $\Pr\left(M_i|y\right) = 1$ and $\Pr\left(M_j|y\right) = 0$. This generalizes Bartlett's 'silly answer'. If we use improper uniform priors for both models then $c_j = O\left(d_j^{n_j}\right)$ and $c_i = O\left(d_i^{n_i}\right)$ and the ratio $c_j/c_i = O\left(d^{n_j - n_i}\right)$ converges or diverges to either 0, 1 or ∞ depending only upon the relative dimensions of the two models. In the first and last cases where $n_i \neq n_j$, 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 such cases the data are unable to inform the posterior probabilities. The fortunate exception, when the sequence $c_j/c_i \to 1$, holds when the dimensions of the models match (see Poirier 1995 and Koop 2003).

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 this conventional wisdom is as follows.

We have two models with parameter vectors θ_i and θ_j which we partition as $\theta_i = (\gamma_i, \overline{\gamma}_i)$ and $\theta_j = \left(\gamma_j, \overline{\gamma}_j\right)$ where $\overline{\gamma}_i$ and $\overline{\gamma}_j$ have the same dimension. If improper priors of the same form are used only on $\overline{\gamma}_k$, $k = i, j$ then we can show that the Bayes factors will be well defined (see for example, Fernández et al., 2001).[2] In this case $c_k = c_{\gamma_k} c_{\overline{\gamma}_k}$ where $c_{\gamma_k} = \int h_k(\gamma_k|\gamma)(d\gamma_k) \leq M < \infty$ and $c_{\overline{\gamma}_k} = \int g(\overline{\gamma}_k) d\overline{\gamma}_k = \infty$ thus $c_j/c_i = c_{\gamma_j}/c_{\gamma_i}$ since the $c_{\overline{\gamma}_i}/c_{\overline{\gamma}_j} \to 1$ at all points in the sequence.

A common example of where this result is used is when an improper prior is placed upon the variance or intercept of the error in a regression model. However, it is not necessary that the $\overline{\gamma}_k$ have the same interpretation under both models. It is necessary that $\overline{\gamma}_k$ have priors that diverge at the same rate in both models. For example we could have the two models

$$y_i = x_i\beta + \sigma\varepsilon_i \text{ and } y_i = z_i\gamma + \sigma\varepsilon_i$$

and the priors for $\beta \in B \subseteq R^{k_i}$ and $\gamma \in \Gamma \subseteq R^{k_j}$ are both divergent in $d_i = diam(B)$ and $d_j = diam(\Gamma)$ such that these priors are $O\left(d_i^{k_i}\right)$ and $O\left(d_j^{k_j}\right)$ respectively. If the priors are such that $O\left(d_i^{k_i}\right) = O\left(d_j^{k_j}\right)$ then the Bayes factor will be well defined. For example, if the prior measures on B and Γ are both uniform, then the condition $O\left(d_i^{k_i}\right) = O\left(d_j^{k_j}\right)$ would imply that the vectors β and γ have the same dimension,

[2] Of course the prior for the entire vector 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. This seems to be accepted language.

$k = k_i = k_j$.

The above result could be thought of as the basis of the next part of this paper where we reparameterize to isolate a parameter, the norm of θ, with common support with divergent prior. In fact, it is not necessary that supports for the norms be the same. Rather they need only be unbounded above some finite value for each model and this value need not be the same for any two models. Related discussion of this issue can be found in, for example, Bartlett (1957), Zellner (1971), O'Hagan (1995), Berger and Perrichi (1996) and Lindley (1997). The above discussion leads us to a number of results which we demonstrate in the remainder of the paper.

For the results that follow, we need to decompose the measure c_i into its different components as they require different treatment. We introduce the necessary concepts and notation here. 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 $v \in V_{1,n}$. Any $n \times 1$ vector $\theta \in A \subseteq R^n$ can be decomposed as $\theta = v\tau$ where $\nu \in V_{1,n}$ which defines the direction of θ and $\tau \in T \subseteq R^+$ defines the vector length. The compact space $V_{1,n}$ has a measure dv_1^n and finite volume

$$\varpi_n = \int_{V_{1,n}} dv_1^n = 2\pi^{n/2}/\Gamma\,(n/2) \tag{1}$$

(Muirhead, 1982). We can decompose the differential term for θ into $d\theta = \tau^{n-1}\,(d\tau)\,dv_1^n$.

The expression for the differential term leads to the following explanation for

Bartlett's paradox. We can decompose the integral c into a convergent (finite) part, ϖ_n, and the divergent part, α_n:

$$c = \int_{R^n} d\theta = \int_T \tau^{n-1} (d\tau) \int_{V_{1,n}} dv_1^n = \alpha_n \varpi_n \tag{2}$$

where $T \equiv R^+$,

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

We can treat α_n as the limit of a sequence α_n^d where

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

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 $c_j = \int_{R^{n_j}} d\theta_j = \alpha_{n_j} \varpi_{n_j}$. Recall that the posterior is well defined even if the integral $c_j = \int_{R^{n_j}} h_j(\theta_j) d\theta_j$ 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 $B_{ij} = m_i/m_j = p_i/p_j \times c_j/c_i$ where since $c_i = c_j = \alpha_n \varpi_n$, $B_{ij} = p_i/p_j$ is well defined since by (4) $c_j/c_i = 1$. 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.

As a simple example to demonstrate this point, consider the uniform distribution on R^n for an n vector of parameters in one model, and a uniform distribution on the same space for an n vector of parameters in another model. If we restrict the supports to have diameter d then the prior measures for the two models become

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

and

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

We will use variants of the rather simple result[3]

$$\lim_{d\to\infty} \frac{\alpha_n^d}{\beta_n^d} = \lim_{d\to\infty} \frac{\int_0^d \tau^{n-1} d\tau}{\int_0^d \nu^{n-1} d\nu} = \lim_{d\to\infty} \frac{nd^n}{nd^n} = 1. \tag{4}$$

Further where $q > 0$

$$\lim_{d\to\infty} \frac{\alpha_{n+q}^d}{\alpha_n^d} = \infty. \tag{5}$$

Despite the apparent simplicity of these results, some of their implications for model comparison with improper priors seem to have been overlooked.

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)

[3]In this simple case, the Theorem 1 can be proved using l'Hopital's rule.

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 (4) and (5) 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 (5) 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

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

However, the Bayes factor B_{ij} is again undefined.

3 Obtaining well defined Bayes factors with divergent priors

The problematic component of the Bayes factor is the ratio c_j/c_i. 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. Other approaches include: Spiegelhalter and Smith (1982); Klein and Brown (1984); O'Hagan (1995); Berger and Pericchi (1996); Phillips and Ploberger (1996); Phillips (1996); and Kleibergen (2004). The

most relevant approach for this paper is that of Klein and Brown (1984, hereafter KB). KB only consider the normal linear regression model while we do not place constraints on the model set. However, both this paper and KB use a sequence of proper priors that, at the limit, are improper and the approach taken by KB to minimizing information most closely resembles the approach taken in this paper.

For a fixed sample size in model i, KB take zero mean Normal priors for regression coefficients θ_i with precision matrix $\underline{V}_i$. For model j denote the precision matrix by $\underline{V}_j$. The approach of KB assumes an unbounded support, but as a proper prior is used the parameters are bounded in probability. KB use a sequence of priors such that these probability bounds increase as $\underline{V}_i \to 0$ and $\underline{V}_j \to 0$ at such rates that $c_j/c_i \to 1$. The sequence is determined by the rate at which $\underline{V}_i \to 0$, and this rate in turn depends upon the dimension of θ_i. KB use limits of measures of information based upon those developed by Shannon (1948) to formalize the concept of 'minimizing information'. Interestingly, for the particular model and prior they consider, they obtain the same expression as Schwarz to approximate the posterior odds ratio.

Our approach is similar to that of KB in that we use a sequence of priors to control c_j/c_i, but this ratio need not converge to one. We use bounded supports but it is the rate of increase of these bounds that governs c_j/c_i and, as in KB, this rate is determined by the dimensions of the models.

We set c_j/c_i by the appropriate choice of the support diameters, d_i and d_j. One

option, which we use in the application, is to assume $\frac{c_j}{c_i} = 1$, such that this ratio plays no further role in the model selection or comparison. Alternatively, one might prefer a term that introduces a penalty for the dimension of the model with a smooth increase in the measure as n increases, but which results in a well defined term in the Bayes factor that does not give unmitigated support for the smallest (or largest) model.

In the simple cases we have considered, the normalizing constant c_i for the improper prior diverged at a rate governed by d_i and n_i. This is not always the case and different measures will diverge at different rates depending upon model dimension. Provided the priors diverge in the support diameters, we may choose the relative diameters ($d = d_j/d_i$) by a rule such that the ratio c_j/c_i is a (possibly constant) function only of the relative dimensions n_j and n_i ($n = n_j/n_i$). This rule may provide a sensible penalty for model dimension if desired.

To demonstrate this idea, consider the Uniform measure on a spherical support centered at the origin and of diameter d_i. In this case $c_i = \frac{\varpi_{n_i} d_i^{n_i}}{n_i}$. Say we choose d_i by the rule $c_i = \delta_0 \delta^{\frac{n_i}{2}} \propto \delta^{\frac{n_i}{2}}$ ($\delta > 0$) such that for all relative diameters d (with sufficiently large d_i) we obtain the Bayes factor $B_{ij} = (p_i/p_j)\,(c_j/c_i) = p_i/p_j \delta^{(n_j - n_i)/2}$. The ratio of prior normalizing constants is

$$\begin{aligned} \frac{c_j}{c_i} &= \delta^{(n_j - n_i)/2} = \frac{n_i \varpi_{n_j} d_j^{n_j}}{n_j \varpi_{n_i} d_i^{n_i}} \\ &= \frac{\varpi_{n_j}}{\varpi_{n_i}} \frac{1}{n} d^{n_j} d_i^{n_j - n_i} \text{ using } d_j = d d_i \text{ and } n_j = n n_i \end{aligned}$$

such that the implied relative diameter d is given by

$$d = \left(\frac{n\varpi_{n_i}}{\varpi_{n_j}}\right)^{1/n_j} \left(\frac{\delta^{1/2}}{d_i}\right)^{(n-1)/n} = \left(n\frac{\Gamma\left(\frac{nn_i}{2}\right)}{\Gamma\left(\frac{n_i}{2}\right)}\right)^{1/nn_i} \left(\frac{1}{d_i}\sqrt{\frac{\delta}{\pi}}\right)^{(n-1)/n}.$$

Choosing $\delta = 1$, such that $c_j/c_i = 1$, simplifies the expression slightly.

For fixed $n > 1$, d increases as δ increases and at a rate determined by n such that larger models have larger diameter supports. For small δ, d has a non-monotonic relationship with n for $n_i < 5$ and monotonically increasing for $n_i \geq 4$. These rules only govern the relative diameter d, so the smallest diameter may be chosen large enough such that the above functions are all increasing in δ and n. As the diameter of the support reflects our certainty about the location of the parameter(s), we can regard larger diameters 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 decrease in certainty. These rules do not always imply larger 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 smaller models to allow them to accumulate sufficient volume.

Now we turn to the practical matter of assigning a value to δ. In the application we use $\delta = 1$. Another choice for δ that suggests itself from the literature is $\delta = \pi$

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.

The above rules imply the use of proper priors that allow us to maintain the features of certain improper priors which bring particular benefits to inference such as reducing frequentist risk or invariance (such as with the Jeffreys prior). As this recommendation requires only a decision on the relative dimensions of the supports, or more specifically a choice of value for c_i/c_j, 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 and so will not be practical (or at all necessary) for proper priors: i.e., with convergent measures on unbounded supports.

4 The improper shrinkage prior: An exception to Bartlett's paradox

In this section we establish a class of priors that are an exception to Bartlett's paradox and show an example of an improper prior that results in a well-defined Bayes factor. As has been discussed, many researchers accept that using improper priors on common parameters, such as intercepts and variances, does not result in Bartlett's paradox provided proper priors are used on all other parameters. Here we show that in treating

the norm of the parameter vector as a common parameter (as every vector has a norm), certain improper priors on *all* parameters result in well defined Bayes factors. To give a preliminary explanation, for this class of priors the divergent part of the integral, α_n, diverges at the same rate for all models using this prior such that the ratio $\alpha_{n_j}/\alpha_{n_i}$ is finite and B_{ij} is well defined. This is effectively using a common form of improper prior on τ.

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, such as estimators using 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 $\|\theta\|^{-(n-2)} = (\theta'\theta)^{-(n-2)/2}$. 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 of all dimensions. The normalizing constant for a model of dimension n is then

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

such that the ratio of the normalizing constants for the Shrinkage priors for models M_i and M_j of different dimensions is always $c_j/c_i = \varpi_{n_j}/\varpi_{n_i}$ given in (6), and this is finite and known.

If we bound the support to diameter d_i, the above normalizing constant becomes $c_i = \frac{\varpi_{n_i} d_i^2}{2}$. Using the rules described in the previous section we can specify $c_i \propto \delta^{\frac{n_i}{2}}$ to obtain

$$\begin{aligned} \frac{c_j}{c_i} &= \delta^{(n_j - n_i)/2} = \frac{\varpi_{n_j} d_j^2}{\varpi_{n_i} d_i^2} = \frac{\varpi_{n_j}}{\varpi_{n_i}} d^2 \\ \text{and } d &= \left(\frac{\varpi_{n_i}}{\varpi_{n_j}}\right)^{1/2} \delta^{(n_j - n_i)/4} = \left(\frac{\Gamma\left(\frac{n n_i}{2}\right)}{\Gamma\left(\frac{n_i}{2}\right)}\right)^{1/2} \left(\frac{\delta}{\pi}\right)^{(n-1)n_i/4}. \end{aligned}$$

From the expression for c_j/c_i above, we can see that the ratio $c_j/c_i = \varpi_{n_j}/\varpi_{n_i}$ could simply result from the choice of relative dimension $d = 1$, and so can hold at all points in the sequence $d_i \to \infty$. In the following section we demonstrate why it is better to choose bounded d_i and set $c_j/c_i = \delta^{(n_j - n_i)/2}$ rather than simply using improper priors and setting $c_j/c_i = \varpi_{n_j}/\varpi_{n_i}$.

5 A pathology and other forms of divergence

In this section we discuss issues related to the computation of Bayes factors with improper priors using the above exception to Bartlett's paradox. First, employing these improper priors for model comparison introduces pathologies into the Bayes factor. These pathologies support our recommendation against using these priors. Second, the case of the Jeffreys prior for the normal linear model is discussed as an example in which the prior does not diverge as the support becomes unbounded; rather it diverges as the complement of the support shrinks. This case may be treated in much the same way as the first case.

5.1 The role of the prior measure

The rules proposed earlier for choosing the relative diameter d set c_j/c_i to a specific value. However, we saw for the uniform and shrinkage priors that this ratio was made up of the ratio $\varpi_{n_j}/\varpi_{n_i}$ and a function of the diameters d_i and d_j. As $\varpi_{n_j}/\varpi_{n_i}$ is always a finite constant it would seem unnecessary to include this in the function for d. This subsection discusses the practical implications of ignoring the term $\varpi_{n_j}/\varpi_{n_i}$ in deciding the diameters. We present a pathology associated with such an approach that suggests we should not in fact refer to the Bayes factors for the unbounded shrinkage prior as 'well defined', but rather as 'able to be calculated'. The main problem is that an important function of the prior measure, penalizing large models,

is lost with this case.

With many proper priors the ratio c_j/c_i brings into the posterior analysis penalties for greater model dimension and greater prior parameter uncertainty. With the Shrinkage prior, the penalty for uncertainty reflected in the support diameter is removed (effectively matched for each model). As shown in the previous section, with unbounded support the ratio is only a function of the dimensions of the models via the ratio $c_j/c_i = \varpi_{n_j}/\varpi_{n_i}$. 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 Assumption 2 held. Further, 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 $\varpi_{n_j}/\varpi_{n_i}$. That is we could specify

$$B_{ij} = \frac{p_i}{p_j} \frac{\varpi_{n_j}}{\varpi_{n_i}}.$$

Although the ratios of prior normalizing constants are identical in each of these examples, equal to $\varpi_{n_j}/\varpi_{n_i}$, they do not produce the same Bayes factor as the ratios p_i/p_j will differ, however they provide useful comparisons for discussion.

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 penalized in the posterior via the prior measure. Because of the behaviour of the ϖ_n over n, the penalty for dimension with the priors that result in the ratio $c_j/c_i = \varpi_{n_j}/\varpi_{n_i}$ 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 $c_j/c_i = \varpi_{n_j}/\varpi_{n_i}$ 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 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) that very small models (small n) are given less penalty than slightly larger models (small $g > 1$) and 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, the

exception being the shrinkage prior.

5.2 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 is possible to ignore the term involving the unit vector v only because the priors discussed 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

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

(see the Appendix for the results in this section). 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[4]. Thus adaptations of priors that result in polynomials in the norm of matching order will

[4]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.

not remove this divergence.

The effect of the divergence in $\widetilde{\varpi}_k^n$ could be removed and Bayes factors computed if the elements of the unit vector v for the variances, the v_{ii}, were restricted 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.

Similarly to the approach discussed in the previous section in which no specific upper bound was placed upon the support diameter, in this case no specific lower bound c_i needs to be specified as it is the relative size of the lower bound that matters. That is, the prior diverges as the area around the origin that is excluded from the support shrinks. Therefore it is the diameter of this complement that governs the rate of divergence and can be chosen to ensure well defined and well behaved Bayes factors.

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 how ν enters the prior.

6 Monte Carlo study

In this section we conduct a small Monte Carlo study to consider computation of the Bayes factors with uniform, shrinkage and Jeffreys priors for a small range of models. We are interested in the ability of these priors to select the appropriate model. The data generating process (DGP) is always the following:

$$\begin{aligned} \rho(L)\, y_t &= \pi(L)\, x_t + \varepsilon_{y,t} \quad \varepsilon_{y,t} \sim N\left(0, \sigma_y^2\right) \\ \Delta x_t &= \mu x_{t-1} + \varepsilon_{x,t} \quad \varepsilon_{x,t} \sim N\left(0, \sigma_x^2\right) \end{aligned}$$

where $E\left(\varepsilon_{y,t}\varepsilon_{x,t}\right) = 0$, $\rho(L) = 1 - 0.7L - 0.15L^2$, $\pi(L) = 0.35 - 0.25L + 0.05L^2$ such that $\rho(1) = \pi(1)$, and $\mu \in \{-0.50, -0.45, \ldots, -0.05, 0\}$.

We include the above DGP in our model set but assume all parameter values are unknown although we know $\rho(1) = \pi(1)$. Denote this model by M_1. The study includes three other models that are similar, in that they capture the essential features of the DGP, but may be over parameterized or incorrectly specified in some respect. There is little reason to explore models that differ greatly from the true DGP as they are easily rejected as our results show.

The second model, M_2, is identical to the DGP above but we do not impose the restriction $\rho(1) = \pi(1)$. The third model, M_3, is a VAR but in the Beveridge Nelson decomposition form. That is, for the vector $z_t = (y_t \quad x_t)$ we estimate

$$\Delta z_t = \Pi z_{t-1} + \Gamma \Delta z_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim N(0, \Sigma).$$

This model differs from the DGP in subtle but important ways. It is in all cases over-parameterized having four more parameters than M_1 and three more than M_2. The extra lags of x_t and y_t in the equation for x_t allow for much richer dynamics in the system, particularly when $\mu = 0$. At the point $\mu = 0$ the matrix Π should have reduced rank which would approximate the final model.

The final model, M_4, is a VECM with known error correction term $w_t = y_t - x_t$:

$$\Delta z_t = \alpha w_{t-1} + \Gamma \Delta z_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim N(0, \Sigma).$$

This is the only model that is an incorrect specification for the DGP (in the sense that it does not encompass the DGP), although it is incorrect only when $\mu < 0$. When $\mu = 0$ it will be correct but overparameterized as it has two more parameters than the DGP, however it should fit well when μ is in the region near 0. The value of μ has a significant effect upon the ability of the Bayes factors computed under the different priors to select the correct model.

Three priors are used in this study: the flat prior; the shrinkage prior (on the mean equation coefficients); and the Jeffreys prior. Combining the priors with the four

models gives twelve Bayesian models. For each value of μ, the marginal likelihoods are computed for each Bayesian model using Importance sampling (Kloek and van Dijk, 1978) and a Multivariate Student-t distribution with 5 degrees of freedom as the candidate density. In all cases the Bayes factors prefer M_1 with the flat prior. Tables 1 reports the Bayes factor for each model and prior to M_1 with the flat prior. It is clear that the flat prior is strongly preferred for all values of μ. Under the flat prior, M_1 and M_2 are preferred over M_3 and M_4, which suggests the flat prior performs well at selecting the correct model and the model closest to the correct model. The model M_3 is preferred over the more parsimonious but incorrect model M_4 until $\mu \approx -0.10$ when M_4 becomes the preferred model.

The shrinkage and Jeffreys priors produce very different orderings of the models to the flat prior. Models M_3 and M_4 are always preferred to M_1 and M_2, and M_3 is preferred over M_4 until $\mu \approx -0.10$, at which point M_4 becomes the preferred model. These results are somewhat surprising as we expected the Shrinkage prior to prefer more parsimonious models (M_1 and M_2). None of the priors was developed with the intention that they be used for model selection, however these results suggest the flat prior would perform better in this role. The Shrinkage and Jeffreys priors prefer models with richer dynamics and are able to discern when important restrictions (in this case cointegration) become relevant.

Table 1: Bayes factor for each model and prior to M_1 with a flat prior

	Flat Prior			Shrinkage Prior				Jeffreys Prior			
μ	M_2	M_3	M_4	M_1	M_2	M_3	M_4	M_1	M_2	M_3	M_4
-0.50	-3.0	-30.6	-50.0	-1091.8	-1083.6	-901.4	-939.5	-1091.9	-1083.7	-909.0	-945.8
-0.45	-3.1	-30.6	-47.7	-1091.5	-1083.3	-901.4	-934.8	-1092.0	-1083.8	-909.4	-941.4
-0.40	-3.1	-30.7	-45.4	-1090.7	-1082.6	-902.0	-931.0	-1091.6	-1083.5	-910.5	-937.8
-0.35	-3.2	-30.8	-42.9	-1090.9	-1082.7	-901.0	-925.4	-1092.2	-1084.0	-909.9	-932.3
-0.30	-3.3	-30.9	-40.1	-1090.7	-1082.6	-900.7	-919.9	-1092.4	-1084.4	-910.0	-927.0
-0.25	-3.3	-31.0	-37.8	-1091.5	-1083.4	-899.5	-914.3	-1093.4	-1085.6	-909.4	-921.5
-0.20	-3.5	-31.1	-35.3	-1091.2	-1083.2	-899.0	-909.3	-1093.4	-1085.7	-909.3	-916.6
-0.15	-3.6	-31.2	-32.7	-1089.8	-1081.8	-899.8	-905.6	-1092.1	-1084.6	-910.3	-913.0
-0.10	-3.8	-31.5	-30.1	-1090.7	-1082.9	-898.5	-899.5	-1093.2	-1086.0	-909.3	-907.1
-0.05	-4.2	-31.9	-27.2	-1090.0	-1082.5	-898.1	-894.4	-1092.7	-1085.9	-909.1	-902.0
0.00	-5.5	-33.3	-23.4	-1090.1	-1083.6	-896.7	-887.5	-1092.9	-1087.5	-907.7	-895.0

7 Conclusion

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 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, in that they are not known a priori to be zero or infinity. One important example 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 a class of priors that does 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 a method of using the same form as the improper prior distributions but on a compact space - bounded by 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 our preferences; for example they may incorporate a penalty for increased model dimension.

8 Acknowledgements

The authors are grateful to John Geweke, Gael Martin, Chris Sims, Arnold Zellner, and participants at the European Seminar on Bayesian Econometrics 2010 for useful discussion.

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

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.

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

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

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.

Kass, R. E. & 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. *Journal of Econometrics* **123**, 227-258.

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

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

Kloek, T. and H. K. Van Dijk (1978), Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica* **46**, 1-19.

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

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

Lindley, D.V. (1957) A Statistical Paradox. *Biometrika* **44**, 187-192.

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

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.

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

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

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

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

Ni, S. X. & 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. (1994) Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models. *Econometrica*, **62**, 73-93..

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

Phillips, P. C. B. & 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 Ap-*

proach. Cambridge: The MIT Press.

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. & 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.

Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate Normal distribution. In Proceedings of the *Third Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1 Berkeley, CA: University of California Press, 197-206.

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

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

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

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

Zellner, A. & 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.

10 Appendix

Theorem 2 *The exact Jeffreys prior for the multivariate Normal linear regression model has the form*

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

Proof. 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 $|D_m (\Sigma^{-1} \otimes \Sigma^{-1}) D_m| = |D_m^+ (\Sigma \otimes \Sigma) D_m^{+\prime}|^{-1} = 2^{\frac{m(m-1)}{2}} |\Sigma|^{-(m+1)}$ (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 $n = km + \frac{m(m+1)}{2}$. 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

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

Using the form of the Shrinkage prior we have the decomposition

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

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.

11 Figures

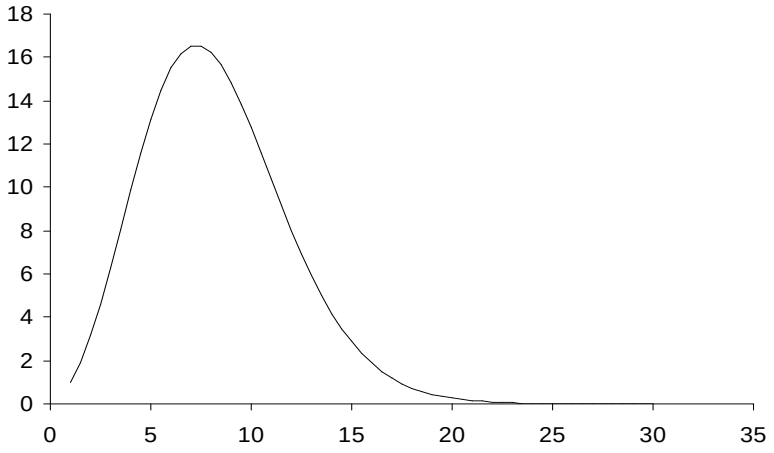

Figure 1: Plot of ϖ_n, the measure for $V_{1,n}$, for $n = 1, ..., 30$.

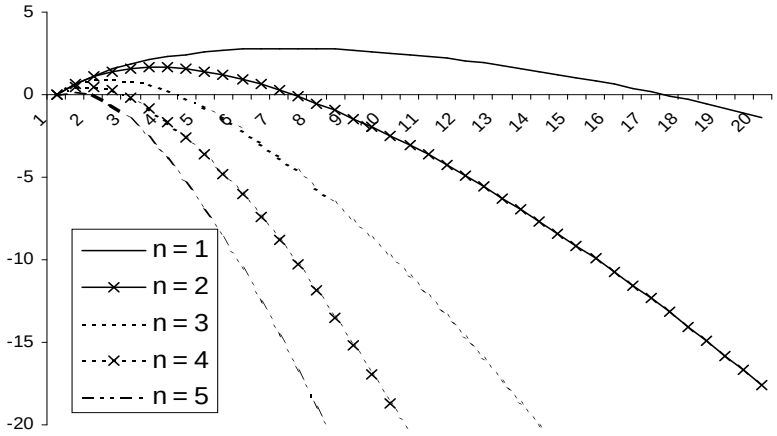

Figure 2: Plot of $\ln(\varpi_{gn}) - \ln(\varpi_n)$ for $n = 1, 2, 3, 4$ and 5 and $g = 1, ..., 20$. The value g is on the $x - axis$.