Available online at www.sciencedirect.com
SCIENCE DIRECT®
JOURNAL OF Econometrics
Journal of Econometrics 123 (2004) 227–258
www.elsevier.com/locate/econbase

Invariant Bayesian inference in regression models that is robust against the Jeffreys–Lindley's paradox

Frank Kleibergen[a,b,*]

[a] *Department of Economics, Box B, Brown University, Providence, RI 02912, USA*
[b] *Faculty of Economics and Econometrics, Department of Quantitative Economics, University of Amsterdam, Roetersstraat 11, Amsterdam 1018 WB, Netherlands*

Abstract

We obtain the prior and posterior probability of a nested regression model as the Hausdorff-integral of the prior and posterior on the parameters of an encompassing linear regression model over a lower-dimensional set that represents the nested model. The Hausdorff-integral is invariant and therefore avoids the Borel–Kolmogorov paradox. Basing priors and prior probabilities of nested regression models on the prior on the parameters of an encompassing linear regression model reduces the discrepancies between classical and Bayesian inference, like, the Jeffreys–Lindley's paradox. We illustrate the analysis with examples of linear restrictions, i.e. a linear regression model, and non-linear restrictions, i.e. a cointegration and an autoregressive moving average model, on the parameters of an encompassing linear regression model.

JEL classification: C11; C12; C52

Keywords: Conditional densities; Borel–Kolmogorov paradox; Prior and posterior (odds ratios); Cointegration; ARMA

1. Introduction

In Bayesian model comparison, based on Bayes factors and prior and posterior odds ratios, prior probabilities for competing models are assigned independently of the priors on the parameters of these models. When one of the models encompasses the others,

* Corresponding author. Department of Economics, Box B, Brown University, Providence, RI 02912, USA. Tel.: (+1) 401-863-9590; fax: (+1) 401-863-1970.
E-mail address: Frank_Kleibergen@brown.edu (F. Kleibergen).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2003.12.009

the prior on its parameters has a specific value at the points in the parameter space that correspond with the competing nested models. Hence, it can occur that this prior has a low value at the location of a competing nested model while that model has a prior probability equal to the prior probability of the encompassing model. This leads to a distinct difference between classical and Bayesian model comparison which is referred to as the Jeffreys–Lindley's paradox, see e.g. Lindley (1957), Bernardo and Smith (1994), O'Hagan (1994) and Poirier (1995). Similar differences arise when we compare the functional forms of sampling densities of maximum likelihood estimators and posteriors of the parameters, see Kleibergen and Zivot (2003).

The priors and prior probabilities are specified independently because the prior on the parameters of an encompassing model does not imply unambiguous probabilities for the lower-dimensional sub-sets of its parameter space that constitute the nested models. This is known as the Borel–Kolmogorov paradox, see e.g. Kolmogorov (1950), Drèze and Richard (1983), Billingsley (1986) and Wolpert (1995). The paradox results because distinct sequences of sets in the parameter space of the encompassing model can be constructed, that converge to a lower-dimensional set and are needed to obtain its probability, which, however, lead to different expressions of the probability of the lower-dimensional set. To avoid the Borel–Kolmogorov paradox, we show that it is necessary to use limit sequences that converge uniformly, i.e. that converge appropriately for all possible sub-sets of the lower-dimensional set. All uniformly converging limit sequences lead to the same expression of the Hausdorff-integral over the lower-dimensional set. This expression implies the prior or posterior probability of a nested model that is induced by the prior or posterior on the parameters of an encompassing model. Also, the priors and posteriors on the parameters of nested models result from these Hausdorff-integrals.

The discrepancies between classical and Bayesian model comparison reduce when we use priors and prior probabilities for nested regression models that result from a prior on the parameters of an encompassing linear regression (ELR) model. Depending on the specification of the ELR model, the Jeffreys–Lindley's paradox is partially or fully overcome. The Jeffreys–Lindley's paradox is fully overcome when the classical "t-values" are the parameters of interest in the ELR model. Just the sensitivity with respect to the prior variance is overcome when we use the parameters of the ELR model as the parameters of interest. Similar results hold for the functional forms of the posteriors and sampling densities of maximum likelihood estimators which are also more in line with one another when we use priors that are induced by the prior on the parameters of an ELR model.

The paper is organized as follows. In the second section, we motivate our analysis and specify the regression models that are nested in the ELR model. In the third section, we discuss the uniformly converging limit sequence by which we avoid the Borel–Kolmogorov paradox and that allows us to obtain the Hausdorff-measure of a lower-dimensional set. In the fourth section, we obtain the prior and prior probability for each nested regression model that is induced by the prior on the parameters of the ELR model. We introduce an example, that we use throughout the paper, to illustrate what this prior probability amounts to. The fifth section extends the results to the posterior and posterior probability. In the sixth section, we discuss the

Jeffreys–Lindley's paradox. We use our example to show, that depending on the parameter of interest in the ELR model, that the Jeffreys–Lindley's paradox is partially or fully overcome. These robustness properties imply that the posterior odds ratio can be constructed even in case of a non-informative prior. The seventh section discusses regression models that are conditional on nuisance parameters. The eighth section contains illustrative examples where we focus on the construction of specifications that satisfy the assumption needed to obtain the uniformly converging limit sequence. We show these for regression models that result from linear restrictions, i.e. linear regression models, and non-linear restrictions, the cointegration and autoregressive moving average (ARMA) model, on the parameters of an ELR model. Finally, the ninth section concludes.

We use the following notation throughout the paper: $\text{vec}(A)$ stands for the column vectorization of the $T \times k$ matrix A such that $\text{vec}(A) = (a_1' \cdots a_k')'$ when $A = (a_1 \cdots a_k)$. $M_A = I_T - A(A'A)^{-1}A'$, with I_T the $T \times T$ dimensional identity matrix; $J(a,(b,c))$ is the Jacobian of the transformation from a to (b,c) and $|_a$ stands for evaluated in a.

2. Motivation

We consider the ELR model,

$$G : y = X\beta + \varepsilon, \tag{1}$$

with y a $T \times 1$ vector of observations on the dependent variable, X a $T \times k$ matrix that contains the independent explanatory variables, β a $k \times 1$ vector of parameters and ε a $T \times 1$ vector of disturbances. The support of β is the $\mathbb{R}^k$. For expository purposes, we assume that $\varepsilon \sim \text{N}(0, I_T)$. This distributional assumption can however be generalized which we further discuss in Section 7. We specify a prior on β in model G, $p_G(\beta)$, that is continuous and continuous differentiable.

We compare and analyze the regression models

$$G_i : y = Xf_i(\varphi_i) + \varepsilon, \quad i = 1, \ldots, n \tag{2}$$

with $\varphi_i \in \Theta_{G_i}$, Θ_{G_i} is an open convex set in the $\mathbb{R}^{m_i}$ and f_i is a k-dimensional continuous differentiable function of the $m_i \times 1$ vector φ_i, $m_i \leqslant k$. Each model G_i in (2) is nested in the ELR model G in (1).

Traditionally, a prior is specified on φ_i in G_i, $p_{G_i}(\varphi_i)$, without considering that G_i is nested in G. This implies that the restrictions that G_i implies on G are directly imposed and that the transformation from G_i to G is left aside, see e.g. Drèze and Richard (1983). The behavior of the posterior of φ_i in G_i can then be pathological when $f_i(\varphi_i)$ is a non-linear function while the posterior of β in G is well-behaved. Examples of such models G_i are cointegration and simultaneous equations models, see Kleibergen and van Dijk (1994, 1998) and Kleibergen and Zivot (2003). Similarly, the prior probability for G_i, $\Pr[G_i]$, is specified without taking the prior on β in G into account. It can therefore occur that the prior on β has a relatively low value at the parameter values of β that correspond with G_i while the prior probability of G_i is relatively large. To overcome these controversies, we propose a framework which explicitly takes into account that G_i is nested in G for the construction of the prior for φ_i and the prior probability of G_i. We therefore base both the prior for φ_i and the prior probability of G_i on the prior on β in G, $p_G(\beta)$. Since the likelihood is a continuous function of the parameters, the results directly extend to the posterior and posterior probabilities.

3. Hausdorff-integrals over lower-dimensional sets

We represent the nested regression models G_i in (2) by lower-dimensional sets in the parameter space of β, the $\mathbb{R}^k$,

$$S_{G_i} = \{\varphi_i \in \Theta_{G_i} \subset \mathbb{R}^{m_i} | \beta = f_i(\varphi_i)\}, \quad i = 1, \ldots, n. \tag{3}$$

The sets S_{G_i} in (3) are m_i-dimensional manifolds in the $\mathbb{R}^k$. We use the sets S_{G_i} to induce the prior probability and density of the regression models G_i from the prior $p_G(\beta)$ on β in G.

The prior $p_G(\beta)$ induces prior probabilities for convex k-dimensional sets $S \subset \mathbb{R}^k$,

$$\Pr_G[S] = \int_S p_G(\beta)\,\mathrm{d}\beta, \tag{4}$$

where $\mathrm{d}\beta$ is shorthand notation for $L_k(\mathrm{d}\beta)$ because (4) is a Lebesgue-integral. To construct the prior probability of G_i induced by $p_G(\beta)$, we evaluate the integral of $p_G(\beta)$ over S_{G_i}. When S in (4) is an m-dimensional manifold and m is less than k, the Lebesgue-measure of S in $\mathbb{R}^k$, $L_k(S)$, equals zero. Hence, we cannot use standard Lebesgue integration to obtain the integral of $p_G(\beta)$ over S. Instead of the Lebesgue-measure, we therefore use a measure that is defined for lower-dimensional sets: the Hausdorff-measure, see e.g. Billingsley (1986) and Rogers (1999).

Definition 1. The Hausdorff-measure of an m-dimensional set S in the $\mathbb{R}^k$, $H_m(S)$, $k \geqslant m$, is the infimum of positive numbers y such that for every $r > 0$, S can be covered by a countable family of closed sets, each of diameter less than r and the sum of the mth powers of these diameters is less than y:

$$H_m(S) = \inf c_m \sum_j \mathrm{diam}(B_j)^m \tag{5}$$

with B_j, $j = 1, \ldots$ the countable number of sets whose union covers S, $\mathrm{diam}(B_j)$ stands for the diameter of B_j, $\mathrm{diam}(B_j) = \sup[|a - b| : a, b \in B_j]$, and c_m is a normalizing constant. The Hausdorff-measure is invariant to the specification of S and can be infinite.

When m equals k, the definition of the Hausdorff-measure in (5) gives the volume of S and the Hausdorff-measure and the Lebesgue-measure coincide if c_k is specified appropriately. To obtain the Hausdorff-measure of the m_i-dimensional set S_{G_i}, where m_i is less than k, we use an invertible mapping of β, that spans the $\mathbb{R}^k$, into φ_i and an additional $(k - m_i)$-dimensional parameter vector λ_i that is such that when λ_i converges to zero $\beta(\varphi_i, \lambda_i)$ converges uniformly to $f_i(\varphi_i)$, i.e. for all values of φ_i.

Assumption 1. In model G from (1), the $k \times 1$ dimensional vector β is an invertible function of the $m_i \times 1$ dimensional vector φ_i and the $(k-m_i)\times 1$ dimensional vector λ_i:

$$\beta = f_i(\varphi_i) + g_i(\varphi_i, \lambda_i), \tag{6}$$

where $g_i(\varphi_i, \lambda_i)$ is a continuous differentiable $k \times 1$ vector function of (φ_i, λ_i) which is such that:

(a) $g_i(\varphi_i, \lambda_i) = 0 \Leftrightarrow \lambda_i = 0$.
(b) The set of values of φ_i that lead to a unique value of $f_i(\varphi_i)$, or for which $\partial f_i/\partial \varphi_i'$ has full rank, is identical to the set of values of φ_i that lead to a unique value of $f_i(\varphi_i) + g_i(\varphi_i, \lambda_i)$, or for which $((\partial f_i/\partial \varphi_i') + (\partial g_i/\partial \varphi_i') \vdots \partial g_i/\partial \lambda_i')$ has full rank, and the latter set does not depend on λ_i, such that $g_i(\varphi_i, \lambda_i)$ is a strictly monotonic function of λ_i for all values of φ_i.[1]
(c) $(\partial g_i(\varphi_i, \lambda_i)/\partial \lambda_i')'(\partial g_i(\varphi_i, \lambda_i)/\partial \lambda_i') \equiv A_i$ for all values of (φ_i, λ_i), with A_i a fixed positive definite symmetric $(k - m_i) \times (k - m_i)$ matrix that does not depend on (φ_i, λ_i).

Definition 1 implies that the Hausdorff-measure of S_{G_i} results from a countable number of sets whose union covers S_{G_i}. Definition 1 does not lead to a straightforward manner of constructing the Hausdorff-measure because such a union of sets can be difficult to obtain. Assumption 1 alleviates the construction of the Hausdorff-measure by specifying a functional relationship between β and (φ_i, λ_i). We can use this functional relationship in two different manners to obtain the Hausdorff-measure of S_{G_i}.

Assumption 1 allows us to project β onto $f_i(\varphi_i)$ so all sets $\beta(\Theta_{G_i}, \Lambda_{i_j})$, with Λ_{i_j} an open convex set in the $\mathbb{R}^{k-m_i}$, are projected onto S_{G_i}. A Hausdorff-measure of S_{G_i} can then be constructed by integrating a distance-measure of λ_i with respect to an integrable function of λ_i, like, for example, a density function, see e.g. McCulloch and Rossi (1992) and Doster (1998). This leads to the Hausdorff-measure of S_{G_i} that is marginal with respect to λ_i since it results from integrating over λ_i.

We construct the Hausdorff-measure of S_{G_i} in the alternative conditional manner that results from conditioning on $\lambda_i = 0$. The Hausdorff-measure is then obtained through a sequence of sets in $\mathbb{R}^k$ that monotonically and uniformly converges to S_{G_i}. Assumption 1(a) implies that λ_i reflects the difference between β and $f_i(\varphi_i)$. Assumption 1(b) is a technical condition that ensures that λ_i reflects this difference for all points φ_i where $\partial f_i/\partial \varphi_i'$ has full rank. Assumption 1(b) is just a translation of the necessary condition of uniform convergence, i.e. convergence for all values of φ_i for which $\partial f_i/\partial \varphi_i'$ has full rank. Assumption 1(c) implies that the difference between β and $f_i(\varphi_i)$ does not depend on φ_i. The normalizing constant that is used for the Hausdorff-measure of sub-sets of S_{G_i} is then identical for all sub-sets of S_{G_i}. This is necessary for an appropriately defined normalizing constant. Assumption 1 is thus a necessary and sufficient condition

[1] We note that this condition refers to the functional relationship $f_i(\varphi_i) + g_i(\varphi_i, \lambda_i)$. The spaces where φ_i, λ_i result from are therefore considered unrestricted, $\varphi_i \in \mathbb{R}^{m_i}$, $\lambda_i \in \mathbb{R}^{k-m_i}$, such that Θ_{G_i} is not involved and, for example, the intersection of the set of values of φ_i that do not imply a unique value for $f_i(\varphi_i)$ and Θ_{G_i} can even be empty.

to obtain an invariant expression of the Hausdorff-measure that results from a limit sequence of sets that converges to the lower-dimensional set.

Assumptions 1(a)–(b) enable the construction of a limit sequence of sets that converges monotonically and uniformly to a sub-set of S_{G_i}. To construct the Hausdorff-measure of such a m_i-dimensional set W_{G_i} ($\subset S_{G_i}$) in the $\mathbb{R}^k$,

$$W_{G_i} = \{\varphi_i \in \Omega_{G_i} \subset \Theta_{G_i} | \beta = f_i(\varphi_i)\}, \quad i = 1, \ldots, n \tag{7}$$

with Ω_{G_i} a convex open m_i-dimensional sub-set of Θ_{G_i}, we use the k-dimensional set

$$W_{G_i}(\rho) = \{\varphi_i \in \Omega_{G_i},\ \lambda_i \in B_{k-m_i}(0,\rho) \subset \mathbb{R}^{k-m_i} | \beta = f_i(\varphi_i) + g_i(\varphi_i, \lambda_i)\}, \tag{8}$$

where $B_{k-m_i}(0,\rho)$ is a $(k-m_i)$-dimensional sphere with radius ρ and the $(k-m_i)$-dimensional vector of zeros as its center. The set $W_{G_i}(\rho)$ contains W_{G_i} for all values of ρ and for a sequence of values of ρ, $\rho_1 > \rho_2 > \cdots > \rho_n > 0$, Assumptions 1(a)–(b) imply that

$$W_{G_i}(\rho_n) \subset W_{G_i}(\rho_{n-1}) \subset \cdots \subset W_{G_i}(\rho_2) \subset W_{G_i}(\rho_1) \quad \text{for all convex sets } \Omega_{G_i} \subset \mathbb{R}^{m_i}. \tag{9}$$

This implies that the convergence of $W_{G_i}(\rho_n)$ to W_{G_i} is strictly monotonic and holds for all sets Ω_{G_i},

$$\lim_{\rho \to 0} W_{G_i}(\rho) = W_{G_i} \quad \text{for all convex sets } \Omega_{G_i} \subset \mathbb{R}^{m_i}. \tag{10}$$

To obtain the Hausdorff-measure of W_{G_i}, we use the Lebesgue-measure of $W_{G_i}(\rho)$ which we normalize to account for the difference in dimension between the m_i-dimensional set W_{G_i} and the k-dimensional set $W_{G_i}(\rho)$. The normalizing constant is the inverse of the Lebesgue-measure of the transformation of the sphere $B_{k-m_i}(0,\rho)$ by the function g_i, see Appendix A for a proof:

$$c_i(\rho)^{-1} = |A_i|^{1/2} V_{k-m_i}(\rho) \tag{11}$$

with $V_{k-m_i}(\rho)$ the volume of a $(k-m_i)$-dimensional sphere with radius ρ. Assumption 1(c) ensures that the normalizing constant does not depend on (φ_i, λ_i) and is therefore the same for every sub-set W_{G_i} of S_{G_i}. The Hausdorff-measure of W_{G_i} then results as the limit when ρ converges to zero of the product of the normalizing constant $c_i(\rho)$ and the Lebesgue-measure of $W_{G_i}(\rho)$:

$$H_{m_i}(W_{G_i}) = \lim_{\rho \to 0} c_i(\rho) L_k(W_{G_i}(\rho)). \tag{12}$$

The normalizing constant $c_i(\rho)$ converges to infinity when ρ converges to zero in such manner that it offsets the convergence to zero of the Lebesgue-measure of $W_{G_i}(\rho)$. The Hausdorff-measure of W_{G_i} in (12), which is a transformation of Ω_{G_i} by f_i, results from conditioning on a zero value of λ_i and does not result from integrating over λ_i.

Theorem 1. *When m_i is less than k and Assumption 1 holds, the Hausdorff measure $H_{m_i}(W_{G_i})$ in (12) is equal to*

$$H_{m_i}(W_{G_i}) = \int_{\Omega_{G_i}} \left| \left(\frac{\partial f_i}{\partial \varphi_i'}\right)' M_{\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)} \left(\frac{\partial f_i}{\partial \varphi_i'}\right) \right|^{1/2} \mathrm{d}\varphi_i \tag{13}$$

and is invariant with respect to transformations of β *and* (φ_i, λ_i) *that satisfy Assumption* 1 *and control for the transformation of* β.

Proof. See Appendix A. □

The definition of the Hausdorff-measure also shows how Hausdorff-integrals of non-negative functions are constructed, see e.g. Billingsley (1986) and Rogers (1999).

Definition 2. When m_i is less than k and Assumption 1 holds, the Hausdorff-integral of a non-negative function $q(\beta)$ over the m_i-dimensional set W_{G_i} reads

$$\int_{W_{G_i}} q(\beta) H_{m_i}(\mathrm{d}\beta) = \lim_{\rho \to 0} \left[c_i(\rho) \int_{W_{G_i}(\rho)} q(\beta)\, \mathrm{d}\beta \right]. \tag{14}$$

Theorem 2. *When* m_i *is less than* k *and Assumption* 1 *holds, the Hausdorff-integral of the non-negative function* $q(\beta)$ *over* W_{G_i} *from* (14) *is equal to*

$$\int_{W_{G_i}} q(\beta) H_{m_i}(\mathrm{d}\beta) = \frac{1}{|A_i|^{1/2}} \int_{\Omega_{G_i}} q(\beta(\varphi_i, \lambda_i)|_{\lambda_i=0}) |J(\beta, (\varphi_i, \lambda_i))|_{\lambda_i=0}|\, \mathrm{d}\varphi_i. \tag{15}$$

The Hausdorff-integral in (15) *is invariant with respect to transformations of* β *and* (φ_i, λ_i) *that satisfy Assumption* 1 *and control for the transformation of* β.

Proof. See Appendix A. □

When m_i equals k, the Hausdorff-measure and integral are identical to the Lebesgue-measure and integral. We use the Hausdorff-integrals to construct prior and posterior probabilities and densities of the models G_i that are induced by $p_G(\beta)$.

The Hausdorff-measure in Theorem 1 gives the measure of a lower-dimensional manifold. These lower-dimensional manifolds represent different models that are nested in the ELR model. We use the Hausdorff-measure to construct priors and posteriors of the parameters of these different models. When the functional form of a class of density functions, such as posteriors, is given, a Hausdorff-measure can be used to construct the relative distance between all densities within this class and a specific one of them. These kind of analyzes are conducted in, for example, McCulloch and Rossi (1992) and Doster (1998). McCulloch and Rossi (1992) use projection functions to map β onto $f_i(\varphi_i)$. Doster (1998) uses Hellinger and Kullback–Leibler distance metrics for the involved density functions. Doster (1998) shows that the resulting specification of the Hausdorff-measure corresponds with a Jeffreys prior, i.e. a prior that is proportional to the square root of the determinant of the information matrix. For a specific choice of the parameter β, the latter also holds for the Hausdorff-measure (13) for some nested regression models. When β is the vector of (conditional) classical t-values of the parameters of the ELR model, the Hausdorff-measure from Theorem 1 is identical to the Jeffreys prior in these models. An example of such a model is the instrumental variables regression model, see Kleibergen and Zivot (2003).

Borel–Kolmogorov paradox: The Borel–Kolmogorov paradox, see e.g. Kolmogorov (1950), Drèze and Richard (1983), Billingsley (1986) and Wolpert (1995), implies that the probability of a lower-dimensional set is not unambiguously defined. Theorems 1 and 2 state the Hausdorff-measure and integral for lower-dimensional sets. These integrals lead to probabilities that are invariant with respect to their specification when this specification accords with Assumption 1. Although we cannot solve the Borel–Kolmogorov paradox, Assumption 1 gives thus a manner of specifying probabilities on lower-dimensional sets that avoids it. The Hausdorff-measure and integrals in Theorems 1 and 2 avoid the Borel–Kolmogorov paradox because we use a limit sequence for which a sequence of sets converges uniformly to the restricted set. Uniform convergence is necessary to ensure that the limit sequence always converges to the restricted set. We therefore avoid the issue of non-conglomerability, i.e. an ambiguous way of reflecting the restriction, that is one element of the Borel–Kolmogorov paradox, see e.g. De Finetti (1972). Assumption 1(c) allows us to obtain the normalizing constant for the Hausdorff-measure.

The traditional example of the Borel–Kolmogorov paradox has two random variables, φ and λ, with joint density $p(\varphi, \lambda)$ and wants to condition on a zero value of λ, see e.g. Drèze and Richard (1983) and Wolpert (1995). We can, for example, use either λ or $\mu = \lambda/\varphi$ to construct a limit sequence of sets that converges to the zero value of λ. When we use λ in the limit sequence, $p(\varphi|\lambda)|_{\lambda=0}$ is the density on the restricted set while $|\varphi|\,p(\varphi|\lambda)|_{\lambda=0}$ is the density on the restricted set when we use μ. The difference between these densities reflects the Borel–Kolmogorov paradox and shows De Finetti's (1972) issue of non-conglomerability, i.e. the restriction can be represented in a non-denumerably infinite number of ways. The difference between the densities arises because the limit sequence does not converge uniformly when we use μ to reflect the restriction. If φ equals zero, the limit sequence that involves μ is not defined and hence its convergence is not uniform over all values of φ. Assumption 1(b) implies uniform convergence and the specification that involves μ does thus not satisfy Assumption 1(b). The specification that involves μ does also not satisfy Assumption 1(c). The specification that involves λ satisfies Assumption 1.

We conclude from Assumption 1 that the Borel–Kolmogorov paradox is avoided when we use limit sequences that converge uniformly. A Bayesian specifies the restricted set, i.e. the nested models G_i, a priori and limit sequences that converge uniformly are obvious since they always converge to the sets that reflect the models G_i. This also holds for the traditional example discussed previously. The arisal of the Borel–Kolmogorov paradox when we use other limit sequences further emphasizes this point.

4. Prior density and prior probability

We construct the prior probability of G_i, $i = 1, \ldots, n$, that is induced by $p_G(\beta)$. In order to obtain these probabilities we assume that the set of models G_i, $i = 1, \ldots, n$, is complete.

Assumption 2. The true model is an element of $\{G_i,\ i = 1, \ldots, n\}$ such that the joint prior probability of the regression models G_i, $i = 1, \ldots, n$, is equal to one.

Assumption 2 shows that we consider the models G_i, $i=1,\ldots,N$, as mutually exclusive events, unless they result from functions $f_i(\varphi_i)$ that are invertible transformations of one another, even when one of them equals G and encompasses all the other models. Hence, all sets S_{G_i} constitute a mutually exclusive event, the model G_i, although they are lower-dimensional sets in the $\mathbb{R}^k$. The probabilities for these events result from the Hausdorff-integral over S_{G_i} with respect to the prior $p_G(\beta)$ after an appropriate normalization for the completeness of the set of models G_i, $i=1,\ldots,n$. The Hausdorff-integrals result from Theorem 2.

Theorem 3. *When Assumptions* 1 *and* 2 *hold, the invariant prior probability for model* G_i, $i=1,\ldots,n$, *that is induced by* $p_G(\beta)$ *reads*

$$Pr_G[G_i]=\frac{Q_{G_i}}{Q},\quad i=1,\ldots,n \tag{16}$$

with

$$Q_{G_i}=\int_{S_{G_i}} p_G(\beta)H_{m_i}(\mathrm{d}\beta) \tag{17}$$

and

$$Q=\sum_{j=1}^{w}\int_{\bigcup_{i=1}^{n_j} S_{i_j}} p_G(\beta)H_{m_j}(\mathrm{d}\beta) \tag{18}$$

with w *the number of sets* S_{G_i} *that have a different function* f_i, $w\leqslant n$, n_j *is the number of sets that have the identical function* f_j *(or an invertible transformation thereof),* m_j *is the dimension of* S_{G_j} *and* S_{i_j}, $i_j=1,\ldots,n_j$ *are the sets with the same function* f_j *involved.*

Proof. Follows directly from Theorem 2. The specification of Q ensures the completeness that results from Assumption 2. □

When m_i equals k, the Hausdorff-integral is identical to the Lebesgue-integral and

$$Q_{G_i}=\int_{S_{G_i}} p_G(\beta)\,\mathrm{d}\beta. \tag{19}$$

If m_i is less than k, we obtain from Theorem 2 that

$$Q_{G_i}=\frac{\left[\left.\frac{\partial \Pr_G[\beta(\{\Theta_{G_i},(-\infty,\lambda_i)\})]}{\partial\lambda_i}\right|_{\lambda_i=0}\right]}{\left|\left.\frac{\partial\beta(0,\lambda_i)}{\partial\lambda_i}\right|_{\lambda_i=0}\right|}=\frac{p_G(\lambda_i)|_{\lambda_i=0}}{|A_i|^{1/2}}\left[\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}\,\mathrm{d}\varphi_i\right], \tag{20}$$

where we have used that

$$\begin{aligned}p_G(\varphi_i,\lambda_i)&=p_G(\beta(\varphi_i,\lambda_i))|J(\beta,(\varphi_i,\lambda_i))|\\&=p_G(\varphi_i|\lambda_i)p_G(\lambda_i).\end{aligned} \tag{21}$$

The resulting specification of Q is then

$$Q=\sum_{j=1}^{w-1}\frac{p_G(\lambda_j)|_{\lambda_j=0}}{|A_j|^{1/2}}\left[\int_{\bigcup_{i_j=1}^{n_j}\Theta_{G_{i_j}}} p_G(\varphi_j|\lambda_j)|_{\lambda_j=0}\,\mathrm{d}\varphi_j\right]+\int_{\bigcup_{i=1}^{n_k}S_{G_{i_w}}} p_G(\beta)\,\mathrm{d}\beta$$

with n_k the number of sets of dimension k. Because of Theorem 2, the prior probability (16) is invariant with respect to the specification of β, (φ_i,λ_i) that satisfy Assumption 1.

Example model: For expository purposes, we consider an example with $n=2$. We use this example throughout and explicitly indicate whenever we use it. We use Theorem 3 to obtain the prior probabilities of, respectively, a nested (non-linear) regression model

$$G_1: y=Xf_1(\varphi_1)+\varepsilon \tag{22}$$

with $\varphi_1\in\Theta_{G_1}\subset\mathbb{R}^{m_1}$, such that

$$S_{G_1}=\{\varphi_1\in\Theta_{G_1}\subset\mathbb{R}^{m_1}|\beta=f_1(\varphi_1)\}, \tag{23}$$

where m_1 is less than k and $f_1(\varphi_1)$ continuous and continuous differentiable, and an ELR model,

$$G_2: y=X\beta+\varepsilon \tag{24}$$

with $\beta\in\mathbb{R}^k$ such that $S_{G_2}=\{\beta\in\mathbb{R}^k\}$. Hence, the model set under consideration includes (22) and the encompassing model in (24), that is identical to (1).

The vital element of the applicability of Theorem 3 is the existence of a function $g_1(\varphi_1,\lambda_1)$ which is such that β and (φ_1,λ_1) satisfy the conditions from Assumption 1. It depends on $f_1(\varphi_1)$ whether $g_1(\varphi_1,\lambda_1)$ is straightforward to obtain. We therefore give examples of its specification for some commonly used regression models in Section 8. Alongside Assumption 1, we also make Assumption 2. Because $\int_{S_{G_2}} p_G(\beta)\,\mathrm{d}\beta=1$, we obtain the probabilities induced by $p_G(\beta)$ for S_{G_1} and S_{G_2} from Theorem 3,

$$\Pr_G[S_{G_1}]=\frac{Q_{G_1}}{1+Q_{G_1}},\qquad \Pr_G[S_{G_2}]=1-\Pr_G[S_{G_1}] \tag{25}$$

with

$$Q_{G_1}=\frac{p_G(\lambda_1)|_{\lambda_1=0}}{|A_1|^{1/2}}\left[\int_{\Theta_{G_1}} p_G(\varphi_1|\lambda_1)|_{\lambda_1=0}\,\mathrm{d}\varphi_1\right]. \tag{26}$$

These prior probabilities imply the prior odds ratio (PROR):

$$\begin{aligned}\mathrm{PROR}[G_1,G_2]&=\frac{\Pr_G[G_1]}{\Pr_G[G_2]}\\ &=Q_{G_1}.\end{aligned} \tag{27}$$

The prior probability from Theorem 3 also implies a prior density of φ_i on Θ_{G_i}.

Theorem 4. *When Assumption* 1 *holds, the prior probabilities* (16) *induce the prior densities*

$$p_{G_i}(\varphi_i) = \lim_{\rho\to 0} \frac{\Pr_G[G_i(\varphi_i,\rho)]}{L_{m_i}[B_{m_i}(\varphi_i,\rho)]}, \qquad i=1,\ldots,n,$$

$$= \frac{p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}}{\int_{\Theta_{G_i}} p_G(u|\lambda_i)|_{\lambda_i=0}\,\mathrm{d}u}, \tag{28}$$

on Θ_{G_i}, *where* $B_{m_i}(\varphi_i,\rho)$ *is an* m_i*-dimensional sphere with radius* ρ *centered at* $\varphi_i\in\Theta_{G_i}$ *and* $Pr_G[G_i(\varphi_i,\rho)]$ *is the prior probability for* G_i *when* φ_i *only results from* $B_{m_i}(\varphi_i,\rho)$. *The prior density* (28) *is invariant with respect to transformations of* β, (φ_i,λ_i) *that satisfy the conditions from Assumption* 1 *and control for the transformation of* β.

Proof. See Appendix A. □

Theorem 4 shows how the prior on the parameters of a nested model is implied by the prior on the parameters of an ELR model. It is often the case, though, that we have a priori ideas about the value of the parameters of the nested model but not of those of the ELR model. The prior on the parameters of the ELR model can, however, be specified such that it accommodates the a priori reasonable values of the parameters of the nested model. For instance, if $\varphi_{i,0}$ is an a priori plausible value of φ_i, we could specify a normal prior on β with mean $f_i(\varphi_{i,0})$. The variance of the prior on β can be used further to accommodate ones a priori thoughts. In Kleibergen and Paap (2002), the prior on the parameters of an ELR model that encompasses a cointegration model is specified in this manner. Reasonable parameter values in the cointegration model are used to specify the prior on the parameters of the ELR model.

Example model: Theorem 4 implies the prior on φ_1 in G_1:

$$p_{G_1}(\varphi_1) = \frac{p_G(\varphi_1|\lambda_1)|_{\lambda_1=0}}{\int_{\Theta_{G_1}} p_G(u|\lambda_1)|_{\lambda_1=0}\,\mathrm{d}u} \tag{29}$$

and on β in G_2:

$$p_{G_2}(\beta) = p_G(\beta). \tag{30}$$

Theorem 4 implies that these priors are invariant with respect to the specification of (φ_1,λ_1) and β that satisfy Assumption 1.

Theorems 3 and 4 show how we conduct Bayesian inference in regression models that are non-linear in the parameters in a manner that is consistent with the Bayesian inference in ELR models. The latter analysis is well-developed and Theorems 3 and 4 show how we extend this analysis to regression models that are non-linear in the parameters. For example, sufficient statistics exist for the parameter β in G and we therefore know how the prior influences the posterior, see e.g. Box and Tiao (1992) and Chao and Phillips (1998). By specifying the prior on φ_i in G_i according to Theorem 4, this property also holds for the prior and posterior of φ_i in G_i. We discuss it for the posterior in the next section.

5. Posterior density and posterior probability

The posterior for β in G from (1) is obtained by updating the prior with the likelihood:

$$p_G(\beta|D) = \frac{p_G(\beta)\mathscr{L}(D|\beta)}{\int_{\mathbb{R}^k} p_G(u)\mathscr{L}(D|u)\,\mathrm{d}u}, \tag{31}$$

where $\mathscr{L}(D|\beta)$ is the likelihood function, which in our case of standard normal disturbances corresponds with

$$\mathscr{L}(D|\beta) = (2\pi)^{-(1/2)T} \exp[-\tfrac{1}{2}(y - X\beta)'(y - X\beta)]. \tag{32}$$

However, any other likelihood that is a continuous and continuous differentiable function of β can be used as well. Since the posterior in (31) is a proper density function, and therefore non-negative, we can, analogous to Theorem 3, construct posterior probabilities by usage of Theorem 2.

Theorem 5. *When Assumptions* 1 *and* 2 *hold, the invariant posterior probability for model* G_i, $i = 1, \ldots, n$, *that is induced by* $p_G(\beta|D)$ (31) *reads*

$$Pr_G[G_i|D] = \frac{Q_{G_i|D}}{Q_D}, \quad i = 1, \ldots, n \tag{33}$$

with

$$Q_{G_i|D} = \int_{S_{G_i}} p_G(\beta|D) H_{m_i}(\mathrm{d}\beta) \tag{34}$$

and

$$Q_D = \sum_{j=1}^{w} \int_{\bigcup_{i=1}^{n_j} S_{i_j}} p_G(\beta|D) H_{m_j}(\mathrm{d}\beta). \tag{35}$$

Proof. Results directly from the proofs of Theorem 2. □

When m_i equals k, the Hausdorff-integral is identical to the Lebesgue-integral and

$$Q_{G_i|D} = \int_{S_{G_i}} p_G(\beta|D)\,\mathrm{d}\beta. \tag{36}$$

If m_i is less than k, we use Theorem 2 to obtain that

$$Q_{G_i|D} = \frac{p_G(\lambda_i|D)|_{\lambda_i=0}}{|A_i|^{1/2}} \left[\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i, D)|_{\lambda_i=0}\,\mathrm{d}\varphi_i\right], \tag{37}$$

where

$$\begin{aligned} p_G(\varphi_i, \lambda_i|D) &= p_G(\beta(\varphi_i, \lambda_i)|D)|J(\beta, (\varphi_i, \lambda_i))| \\ &= p_G(\varphi_i|\lambda_i, D)\,p_G(\lambda_i|D). \end{aligned} \tag{38}$$

The accompanying specification of Q_D is given by

$$Q_D = \sum_{j=1}^{w-1} \frac{p_G(\lambda_j|D)|_{\lambda_j=0}\left[\int_{\bigcup_{i_j=1}^{n_j} \Theta_{G_{i_j}}} p_G(\varphi_j|\lambda_j, D)|_{\lambda_j=0}\,\mathrm{d}\varphi_j\right]}{|A_j|^{1/2}} + \int_{\bigcup_{i=1}^{n_k} S_{G_{i_w}}} p_G(\beta|D)\,\mathrm{d}\beta. \tag{39}$$

We refer to Theorem 3 for further clarification of the different symbols. Theorem 2 shows that the posterior probabilities are invariant to the specification of β, (φ_i, λ_i) that satisfy Assumption 1.

Analogous to the result in Theorem 4, the posterior probabilities in (33) also imply a posterior density for φ_i on Θ_{G_i}.

Theorem 6. *When Assumption* 1 *holds, the posterior probabilities* (33) *induce the posterior densities*

$$p_{G_i}(\varphi_i|D) = \frac{p_G(\varphi_i|\lambda_i, D)|_{\lambda_i=0}}{\int_{\Theta_{G_i}} p_G(u|\lambda_i, D)|_{\lambda_i=0}\,\mathrm{d}u}, \quad i = 1, \ldots, n, \tag{40}$$

on Θ_{G_i}, and these posterior densities are invariant with respect to the specification of β, (φ_i, λ_i) that satisfy the conditions from Assumption 1.

Proof. Results directly from the proof of Theorem 4. □

Naturally, the posterior densities (40) also result when we update the prior $p_{G_i}(\varphi_i)$ with the likelihood:

$$p_{G_i}(\varphi_i|D) = \frac{p_{G_i}(\varphi_i)\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}}{\int_{\Theta_{G_i}} p_{G_i}(\psi_i)\mathscr{L}(D|u)|_{u=f_i(\psi_i)}\,\mathrm{d}\psi_i}, \quad i = 1, \ldots, n. \tag{41}$$

Similarly, the posterior probabilities (33) result from the equality between the posterior odds ratio (POR) and the prior odds ratio (PROR) times the Bayes factor (BF):

$$\mathrm{POR}(G_i, G_j) = \mathrm{PROR}(G_i, G_j) \times \mathrm{BF}(G_i, G_j), \tag{42}$$

where

$$\mathrm{POR}(G_i, G_j) = \frac{\Pr_G[G_i|D]}{\Pr_G[G_j|D]}, \quad \mathrm{PROR}(G_i, G_j) = \frac{\Pr_G[G_i]}{\Pr_G[G_j]},$$

$$\mathrm{BF}(G_i, G_j) = \frac{p_{G_i}(D)}{p_{G_j}(D)} \tag{43}$$

and $p_{G_i}(D)$ is the marginal data density,

$$p_{G_i}(D) = \int_{\Theta_{G_i}} p_{G_i}(\varphi_i)\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}\,\mathrm{d}\varphi_i$$

$$= c_\beta \times \frac{p_G(\lambda_i|D)|_{\lambda_i=0}}{p_G(\lambda_i)|_{\lambda_i=0}} \times \frac{\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i, D)|_{\lambda_i=0}\,\mathrm{d}\varphi_i}{\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}\,\mathrm{d}\varphi_i} \tag{44}$$

with $c_\beta^{-1} = \int_{\mathbb{R}^k} p_G(\beta)\mathscr{L}(D|\beta)\,\mathrm{d}\beta$. For a proof of (44) we refer to Appendix A, see also Verdinelli and Wasserman (1995).

The specification of the prior $p_{G_i}(\varphi_i)$ in (28) satisfies the conditions for the Bayes factor to equal the Savage–Dickey density ratio, see e.g. Dickey (1971) and Verdinelli and Wasserman (1995). The Bayes factor is therefore equal to the ratio of the posterior heights divided by the prior heights:

$$\mathrm{BF}(G_i, G_j) = \frac{\left[\frac{p_G(\lambda_i|D)|_{\lambda_i=0}}{p_G(\lambda_i)|_{\lambda_i=0}}\right]\left[\frac{\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i,D)|_{\lambda_i=0}\,\mathrm{d}\varphi_i}{\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}\,\mathrm{d}\varphi_i}\right]}{\left[\frac{p_G(\lambda_j|D)|_{\lambda_j=0}}{p_G(\lambda_j)|_{\lambda_j=0}}\right]\left[\frac{\int_{\Theta_{G_j}} p_G(\varphi_j|\lambda_j,D)|_{\lambda_j=0}\,\mathrm{d}\varphi_j}{\int_{\Theta_{G_j}} p_G(\varphi_j|\lambda_j)|_{\lambda_j=0}\,\mathrm{d}\varphi_j}\right]}. \tag{45}$$

Substituting this expression for the Bayes factor in (42) results in the posterior odds ratio that accords with the one that results directly from the posterior probabilities (33), i.e.

$$\mathrm{POR}(G_i, G_j) = \frac{Q_{G_i|D}}{Q_{G_j|D}}. \tag{46}$$

Example model: The Bayes factor for comparing G_1 with G_2 becomes

$$\mathrm{BF}(G_1, G_2) = \left[\frac{p_G(\lambda_1|D)|_{\lambda_1=0}}{p_G(\lambda_1)|_{\lambda_1=0}}\right]\left[\frac{\int_{\Theta_{G_1}} p_G(\varphi_1|\lambda_1, D)|_{\lambda_1=0}\,\mathrm{d}\varphi_1}{\int_{\Theta_{G_1}} p_G(\varphi_1|\lambda_1)|_{\lambda_1=0}\,\mathrm{d}\varphi_1}\right] \tag{47}$$

and the posterior odds ratio for comparing G_1 and G_2 reads

$$\mathrm{POR}(G_1, G_2) = Q_{G_1|D}$$

$$= \frac{p_G(\lambda_1|D)|_{\lambda_1=0}}{|A_1|^{1/2}}\left[\int_{\Theta_{G_1}} p_G(\varphi_1|\lambda_1, D)|_{\lambda_1=0}\,\mathrm{d}\varphi_1\right]. \tag{48}$$

The first part in the Bayes factor (47) is the Savage–Dickey density ratio, see Dickey (1971) and Verdinelli and Wasserman (1995). The second part arises because the integrals of the conditional densities $p_G(\varphi_1|\lambda_1, D)|_{\lambda_1=0}$ and $p_G(\varphi_1|\lambda_1)|_{\lambda_1=0}$ over Θ_{G_1} do not have to be equal to one. When $\Theta_{G_1} = \mathbb{R}^{m_1}$, the integrals of both conditional densities are equal to one and the Bayes factor simplifies to the usual expression of the Savage–Dickey density ratio.

6. Jeffreys–Lindley's paradox

Example model: To discuss the Jeffreys–Lindley's paradox, we further simplify our example. Model G,

$$G : y = \imath_T \beta + \varepsilon \tag{49}$$

with y a $T \times 1$ vector of observations on the dependent variable and $\imath_T$ a $T \times 1$ vector of ones, now contains only one parameter so β is a scalar and its support is $\mathbb{R}$. We specify a normal prior on β with mean b and variance τ^2.

$$p_G(\beta) = (2\pi\tau^2)^{-1/2} \exp\left[-\frac{1}{2\tau^2}(\beta - b)^2\right]. \tag{50}$$

When we combine the prior with the likelihood, we obtain the posterior,

$$p_G(\beta|D) = (2\pi)^{-1/2}\left[\frac{1}{\tau^2} + T\right]^{1/2} \exp\left[-\frac{1}{2}\left(\frac{1}{\tau^2} + T\right)(\beta - \tilde{b})^2\right] \tag{51}$$

with

$$\tilde{b} = \left(\frac{1}{\tau^2} + T\right)^{-1}\left(\frac{1}{\tau^2} b + T\hat{b}\right) \tag{52}$$

and $\hat{b} = \imath_T' y/T$. G_1 is the model that results when β equals zero,

$$G_1 : y = \varepsilon \tag{53}$$

and G_2 is identical to G. The restriction imposed on β to obtain G_1 is such that it satisfies Assumption 1 for λ identical to β. We use this specification to reflect the difference between G_1 and G_2. The Bayes factor $\mathrm{BF}(G_1, G_2)$ for comparing G_1 and G_2 then equals the Savage–Dickey density ratio (47),

$$\mathrm{BF}(G_1, G_2) = \frac{p_G(\lambda|D)|_{\lambda=0}}{p_G(\lambda)|_{\lambda=0}} = \frac{p_G(\beta|D)|_{\beta=0}}{p_G(\beta)|_{\beta=0}}$$
$$= [1 + \tau^2 T]^{1/2} \exp[-\tfrac{1}{2}(T\tilde{b}^2 + (1/\tau^2)(\tilde{b}^2 - b^2))]. \tag{54}$$

We distinguish two instances of the Jeffreys–Lindley's paradox that imply a degeneracy of the Bayes factor and a distinct difference between classical and Bayesian model comparison, see e.g. Lindley (1957), Bernardo and Smith (1994), O'Hagan (1994) and Poirier (1995):

(a) When the prior variance τ^2 converges to infinity, $\mathrm{BF}(G_1, G_2)$ goes to infinity:

$$\lim_{\tau^2 \to \infty} \mathrm{BF}(G_1, G_2) = \infty. \tag{55}$$

(b) When the number of observations T converges to infinity, $\mathrm{BF}(G_1, G_2)$ goes to zero unless $\tilde{\beta}$ is equal to zero:

$$\lim_{T \to \infty} \mathrm{BF}(G_1, G_2) = 0, \quad \tilde{\beta} \neq 0$$
$$= \infty, \quad \tilde{\beta} = 0. \tag{56}$$

We separately discuss these two instances of the Jeffreys–Lindley's paradox and analyze whether the induced probability approach overcomes either one of these two instances.

(a) The prior odds ratio to compare G_1 and G_2 that results from Theorem 3 and (27) is

$$\begin{aligned}\mathrm{PROR}(G_1,G_2) &= p_G(\lambda(\beta))|_{\lambda=0}|J(\beta,\lambda)|_{\lambda=0}| \\ &= (2\pi\tau^2)^{-1/2}\exp\left[-\frac{1}{2\tau^2}\,b^2\right],\end{aligned} \tag{57}$$

since $A_1=1$. When we multiply the prior odds ratio with the Bayes factor in (54), or make use of Theorem 5, we obtain the posterior odds ratio

$$\begin{aligned}\mathrm{POR}(G_1,G_2) &= p_G(\lambda(\beta))|_{\lambda=0}|J(\beta,\lambda)|_{\lambda=0}| \\ &= (2\pi)^{-1/2}\left[\frac{1}{\tau^2}+T\right]^{1/2}\exp\left[-\frac{1}{2}\left(\frac{1}{\tau^2}+T\right)\tilde{b}^2\right].\end{aligned} \tag{58}$$

The posterior odds ratio in (58) does not converge to infinity when the prior variance τ^2 becomes infinite. Instead, using the expression for $\tilde{b}$ in (52), it can be shown that

$$\lim_{\tau^2\to\infty}\mathrm{POR}(G_1,G_2) = (2\pi)^{-1/2}T^{1/2}\exp[-\tfrac{1}{2}T\hat{b}^2], \tag{59}$$

where the right-hand side of (59) is a finite non-zero constant for finite T. Hence, the posterior odds ratio is well-defined in case of an infinite prior variance. Examples of priors with an infinite variance are non-informative priors. The Bayes factor is infinite when we use such a non-informative prior as shown by (55). Similarly, the prior odds ratio in (57) is equal to zero in case of a non-informative prior, which is obtained by using (57) and letting τ^2 converge to infinity. Theorem 5, however, still gives a well-defined expression for the posterior odds ratio in case of a non-informative prior. The convergence to zero of the prior odds ratio and the convergence to infinity of the Bayes factor therefore cancel each other out in the posterior odds ratio. We can also use Theorem 5 to obtain the prior that leads to the same posterior odds ratio as the limit expression in (59). This non-informative prior is

$$p_G(\beta)\propto 1. \tag{60}$$

When the prior variance converges to infinity, the Bayes factor becomes infinite because of the zero value of the prior in the denominator of the Savage–Dickey density ratio in (54). The finite value of the posterior odds ratio therefore shows that the prior odds ratio offsets the zero value in the denominator of the Bayes factor and thus corrects the Bayes factor for the plausibility of the competing models reflected in the prior.

(b) When T goes to infinity, the posterior odds ratio in (58) converges to

$$\begin{aligned}\lim_{T\to\infty}\mathrm{POR}(G_1,G_2) &= \lim_{T\to\infty}(2\pi)^{-1/2}\left[\frac{1}{\tau^2}+T\right]^{1/2}\exp\left[-\frac{1}{2}\left(\frac{1}{\tau^2}+T\right)\tilde{b}^2\right] \\ &= 0,\quad \tilde{b}\neq 0 \\ &= \infty,\quad \tilde{b}=0.\end{aligned} \tag{61}$$

The convergence behavior of the posterior odds ratio is in this case identical to the convergence behavior of the Bayes factor. The posterior odds ratio equals the product of the prior odds ratio and the Bayes factor. Since the prior odds ratio remains fixed, the posterior odds ratio has the same convergence behavior as the Bayes factor.

Another way to analyze the limit behavior of the Bayes factor and the posterior odds ratio, when T goes to infinity, is to express them using a classical "t-value",

$$\tilde{z} = \tilde{b}\sqrt{\frac{1}{\tau^2} + T}. \tag{62}$$

The expressions of the Bayes factor and the posterior odds ratio then become, respectively,

$$\mathrm{BF}(G_1, G_2) = [1 + \tau^2 T]^{1/2} \exp\left[-\frac{1}{2}\left(\tilde{z}^2 - \frac{1}{\tau^2}\, b^2\right)\right] \tag{63}$$

and

$$\mathrm{POR}(G_1, G_2) = (2\pi)^{-1/2}\left[\frac{1}{\tau^2} + T\right]^{1/2} \exp\left[-\frac{1}{2}\,\tilde{z}^2\right]. \tag{64}$$

When the classical t-value, $\tilde{z}$, remains fixed, both the Bayes factor in (63) and the posterior odds ratio in (64) become infinite when T converges to infinity, see e.g. Berger (1985). Classical statistical analysis has in this case a non-zero probability of rejecting G_1 against G_2 since $\tilde{z}$ remains finite. A Bayesian that uses the posterior odds ratio from (64) or the Bayes factor from (63), however, always chooses G_1.

The classical t-value $\tilde{z}$ is not the sufficient statistic for the posterior of β in (51). To analyze the limit behavior of the Bayes factor in terms of a statistic that is not a sufficient statistic can be considered as rather peculiar. The classical t-statistic is the sufficient statistic of the posterior of ζ with

$$\zeta = \beta\sqrt{\frac{1}{\tau^2} + T}. \tag{65}$$

Expressed in ζ, model G from (49) reads

$$G: y = \imath_T\, \frac{1}{\sqrt{1/\tau^2 + T}}\, \zeta + \varepsilon. \tag{66}$$

Our parameter of interest is β which is identical to $1/\sqrt{1/\tau^2 + T}\zeta$. When expressed using the parameter of interest β, G_1 corresponds with $\beta = 0$. The specification of β as a function of ζ that is in line with Assumption 1 reads

$$\beta = g(\zeta) \tag{67}$$

with

$$g(\zeta) = \frac{1}{\sqrt{1/\tau^2 + T}}\, \zeta.$$

In (67), G_1 is obtained when ζ equals zero.

We construct the posterior odds ratio and Bayes factor for comparing G_1 and G_2 using (67). The prior on β (50) implies the prior on ζ,[2]

$$p_G(\zeta) = (2\pi)^{-1/2}(1+\tau^2 T)^{-1/2}\exp[-\tfrac{1}{2}(1+\tau^2 T)^{-1}(\zeta - z)^2], \tag{68}$$

with $z = b\sqrt{1/\tau^2 + T}$. Similarly, the posterior of ζ that is implied by the posterior of β in (51) reads

$$p_G(\zeta|D) = (2\pi)^{-1/2}\exp[-\tfrac{1}{2}(\zeta - \tilde{z})^2]. \tag{69}$$

The Bayes factor results from the Savage–Dickey density ratio,

$$\begin{aligned}\mathrm{BF}(G_1, G_2) &= \frac{p_G(\lambda(\zeta)|D)|_{\lambda=0}|J(\zeta,\lambda)|_{\lambda=0}|}{p_G(\lambda(\zeta))|_{\lambda=0}|J(\zeta,\lambda)|_{\lambda=0}|}\\ &= [1+\tau^2 T]^{1/2}\exp\left[-\frac{1}{2}\left(T\tilde{b}^2 + \frac{1}{\tau^2}(\tilde{b}^2 - b^2)\right)\right],\end{aligned} \tag{70}$$

which is identical to the expression in (54). The posterior odds ratio that results from the prior odds ratio implied by Theorem 3 and the Bayes factor, or directly from Theorem 5 using (69) and (67), becomes

$$\begin{aligned}\mathrm{POR}(G_1, G_2) &= |J(\beta,\zeta)|_{\zeta=0}|^{-1}p_G(\zeta|D)|_{\zeta=0}\\ &= |J(\beta,\zeta)|_{\zeta=0}|^{-1}(2\pi)^{-1/2}\exp\left[-\frac{1}{2}\tilde{z}^2\right]\\ &= (2\pi)^{-1/2}\left(\frac{1}{\tau^2}+T\right)^{1/2}\exp\left[-\frac{1}{2}\tilde{z}^2\right],\end{aligned} \tag{71}$$

since $A_1 = J(\beta,\zeta)'J(\beta,\zeta)$ such that $|A_1|^{1/2} = |J(\beta,\zeta)|$. The posterior odds ratio in (71) is identical to the posterior odds ratio in (64). This results from the invariance of the Hausdorff-integral stated in Theorem 2. The reason why this invariance occurs is that β is the parameter of interest and we express G_1 in terms of β in both cases.

Instead of β, we can also use ζ as our parameter of interest. We then use (66) as the basic specification of G instead of (49). The posterior of ζ in (69) shows that $\tilde{z}$ is a sufficient statistic for this posterior. A representation of ζ that satisfies Assumption 1 is

$$\zeta = h(\theta) \tag{72}$$

with $h(\theta) = \theta$. G_1 is obtained when θ equals zero. We can also express the posterior odds ratio and Bayes factor using specification (72). The Bayes factor that results from the Savage–Dickey density ratio remains identical to (70). The posterior odds ratio that results from Theorem 5, that equals the posterior from (69) evaluated in $\zeta = 0$ and which can also be obtained from combining the prior odds ratio from Theorem 3 with

[2] We note that the prior on ζ depends on the data, since $T = \iota_T'\iota_T$, and therefore violates the likelihood principle.

the Bayes factor, does, however, change

$$\begin{aligned}\mathrm{POR}(G_1, G_2) &= p_G(\zeta|D)|_{\zeta=0}\\ &= (2\pi)^{-1/2}\exp[-\tfrac{1}{2}\tilde{z}^2].\end{aligned} \tag{73}$$

The difference between the posterior odds ratios (71) and (73) shows that the invariance of the Hausdorff-integral is only with respect to specifications that involve the same parameter of interest. Hence, the posterior odds ratios differ depending on whether β or ζ is the parameter of interest. When we use ζ as our parameter of interest, the posterior odds ratio (73) remains constant when T converges to infinity and $\tilde{z}$ remains fixed. It is also only negligibly affected by an increase in the prior variance τ^2. Hence, when ζ is our parameter of interest neither of the two instances of the Jeffreys–Lindley's paradox occurs.

The posterior odds ratio that we used to compare G_1 and G_2 when β is the parameter of interest, namely (58) (or (64) or (71)), is based on a prior and posterior which are identical to the prior and posterior underlying the posterior odds ratio for ζ in (73). The posterior odds ratios therefore only differ because of the different choice of the parameter of interest or, put differently, the basic specification of G. A Bayesian considers the parameters of a model, from which a set of observations is generated, as a realization from a prior distribution. Hence, for a Bayesian the parameters of the model serve naturally as the parameters of interest. When we use classical t-values as parameters of interest, the resulting Bayesian analysis shares many features with classical statistical tests, like, for example, the robustness to the Jeffreys–Lindley's paradox. So if one wants to conduct Bayesian inference that is closely related to classical statistical analysis, one should use the classical t-values as the parameters of interest. We note, however, that there are several problems attached to such an approach. For example, the priors on the classical t-values violate the likelihood principle and the prior odds ratio therefore depends on the involved data set. Hence, the priors odds ratio changes when the sample size increases. Another problem concerns extensions to a multiple parameter setting where the prior is then on the vector of classical t-values. It is not clear what such a prior implies for each individual parameter.

The Bayes factor is popular for model comparison because it is equal to the posterior odds ratio when the prior odds ratio is equal to one, which implies equal prior probabilities for the competing models, see e.g. Kass and Raftery (1995). This specification of the posterior odds ratio is affected by the Jeffreys–Lindley's paradox. Loosely speaking the Jeffreys–Lindley's paradox implies that the Bayes factor converges to infinity when the prior variance or the number of observations go to infinity, see e.g. Lindley (1957), Bernardo and Smith (1994), O'Hagan (1994) and Poirier (1995). The above example shows that the induced probability approach leads to posterior odds ratios that are partially or fully robust to the Jeffreys–Lindley's paradox. The results from the example extend in a straightforward manner to the general setting of comparing n regression models. It depends on the choice of the parameter of interest whether the posterior odds ratios that result from the posterior probabilities from Theorem 5 are partially or fully robust to the Jeffreys–Lindley's paradox. When we use the parameters of the encompassing model G, β, as the parameters of interest, the posterior odds ratios that result from Theorem 5 are robust against the element of the Jeffreys–Lindley's paradox that is concerned with the prior variance. When we use the classical t-values of the parameters of the encompassing model as the parameters of interest, the posterior odds ratios are robust against both elements of the Jeffreys–Lindley's paradox.

The robustness against the prior variance element of the Jeffreys–Lindley's paradox implies that we can obtain the posterior odds ratio using Theorem 5 in case of an improper prior, like, for example, a non-informative prior. A non-informative prior implies that the prior odds ratio that results from Theorem 3 is equal to zero. The Jeffreys–Lindley's paradox implies that the Bayes factor is infinite when we use a non-informative prior. Theorem 5 can be used whenever the posterior is proper and implies that the zero value of the prior odds ratio cancels out the infinite value of the Bayes factor in case of a non-informative prior. To obtain a proper posterior, it is not necessary that the prior integrates to one.

The posterior odds ratios that result from Theorem 5 are related to the posterior information criterium of Phillips and Ploberger, see e.g. Phillips and Ploberger (1994, 1996) and Phillips (1996). Although the posterior information criterium is intended for time-series only, an identical expression of the posterior odds ratio from Theorem 5 results, apart from the 2π-terms, when the classical t-values are our parameters of interest and we specify a flat prior on them. The $|A_i|^{-1/2}|J(\beta,(\varphi_i,\lambda_i))|$ term in the Jacobian in Theorem 5 is then identical to the Jeffreys' prior which is the penalty term in the posterior information criterium. The posterior information criterium is a ratio of σ-finite posterior measures and does by itself not lead to well-defined posterior odds ratios. It is therefore surprising that when we include the 2π-terms that Theorem 5 indicates that we can interpret the posterior information criterium as a posterior odds ratio.

7. Nuisance parameters

For expository purposes, thus far, we have only discussed regression models that contain no nuisance parameters. When model G in (1) is a linear regression model conditional on a realization of an $l \times 1$ vector of nuisance parameters η, we specify it as

$$G : P_y(\eta)y = P_X(\eta)X\beta + \varepsilon \tag{74}$$

and model G_i as

$$G_i : P_y(\eta)y = P_X(\eta)Xf_i(\varphi_i) + \varepsilon, \quad i = 1,\dots,n, \tag{75}$$

where the $T \times T$ matrices $P_y(\eta)$ and $P_X(\eta)$ are observable given a realization of the nuisance parameter vector η. The matrices $P_y(\eta)$ and $P_X(\eta)$ incorporate the nuisance parameters such that the disturbances $\varepsilon : T \times 1$ have a pre-defined distribution that does not depend on nuisance parameters. We specify a joint prior on (β,η),

$$p_G(\beta,\eta) = p_G(\beta|\eta)p_G(\eta). \tag{76}$$

Theorem 7. *When Assumptions 1 and 2 hold and model G in (74) is a linear regression model given a realization of the nuisance parameter vector η, the expressions of the prior and posterior probabilities in Theorems 3 and 5 induced by $p_G(\beta,\eta)$ and $p_G(\beta,\eta|D)$ remain unaltered when we replace Q_{G_i} and $Q_{G_i|D}$ by*

$$Q_{G_i} = \int_{\Theta_\eta} \left[\int_{\Theta_{G_i}} p_G(\beta|\eta) H_{m_i}(\mathrm{d}\beta) \right] p_G(\eta)\,\mathrm{d}\eta, \quad i = 1,\ldots,n,$$

$$Q_{G_i|D} = \int_{\Theta_\eta} \left[\int_{\Theta_{G_i}} p_G(\beta|\eta,D) H_{m_i}(\mathrm{d}\beta) \right] p_G(\eta|D)\,\mathrm{d}\eta, \quad i = 1,\ldots,n, \tag{77}$$

where Θ_η is the parameter region of η. Similarly, the joint prior and posterior densities of (φ_i,η) defined on $\Theta_{G_i} \times \Theta_\eta$ that result from Theorems 4 and 5 read

$$p_{G_i}(\varphi_i,\eta) = \frac{p_G(\varphi_i,\lambda_i|\eta)|_{\lambda_i=0}\, p_G(\eta)}{\int_{\Theta_\eta} \left[\int_{\Theta_{G_i}} p_G(\varphi_i,\lambda_i|\eta)|_{\lambda_i=0}\,\mathrm{d}\varphi_i \right] p_G(\eta)\,\mathrm{d}\eta}, \quad i = 1,\ldots,n,$$

$$p_{G_i}(\varphi_i,\eta|D) = \frac{p_G(\varphi_i,\lambda_i|\eta,D)|_{\lambda_i=0}\, p_G(\eta|D)}{\int_{\Theta_\eta} \left[\int_{\Theta_{G_i}} p_G(\varphi_i,\lambda_i|\eta,D)|_{\lambda_i=0}\,\mathrm{d}\varphi_i \right] p_G(\eta|D)\,\mathrm{d}\eta}, \quad i = 1,\ldots,n. \tag{78}$$

The probabilities that result from (77) and the densities (78) are invariant with respect to transformations of β, (φ_i,λ_i) that satisfy Assumption 1 and control for the transformation of β.

Proof. Results directly from Theorem 2. □

The prior/posterior probabilities and densities stated in Theorem 7 are invariant with respect to transformations that satisfy the conditions from Assumption 1. They are not invariant to transformations that involve the nuisance parameter η. Invariance to these kind of transformations can be achieved by an appropriate specification of the prior $p_G(\beta,\eta)$ in (76).

8. Examples

We discuss some examples of regression models that result from a restriction on the parameters of an ELR model. For reasons of brevity, we only discuss the construction of the specification that satisfies Assumption 1. The explicit expressions of the priors, prior probabilities, etc. are left aside but are straightforward to construct given the specification that satisfies Assumption 1. The first example concerns linear restrictions that lead to a nested linear regression model. The second and third example are concerned with non-linear restrictions that lead to a cointegration model and an ARMA(1,1) model. Similar results hold for other regression models that are obtained from non-linear restrictions on the parameters of an ELR model, for example, for the instrumental variables regression models, see Kleibergen and Zivot (2003), and the simultaneous equation model, see Kleibergen (1997) and Kleibergen and van Dijk (1998).

8.1. Linear regression model

Our first example considers linear restrictions on the parameters of a linear regression model, see also Tiao et al. (1977),

$$G: y = (X \;\; Z)\beta + u, \tag{79}$$

where $y: T \times 1$, $X: T \times m$, $Z: T \times (k-m)$, m is less than k, $\beta: k \times 1$, $\beta \in \mathbb{R}^k$ and $u \sim \mathrm{N}(0, \sigma^2 I_T)$. Our linear regression model of interest G_1,

$$G_1: y = X\varphi + u \tag{80}$$

with $\varphi: m \times 1$, $\varphi \in \mathbb{R}^m$, is nested in the encompassing model G from (79). We therefore specify S_{G_1} as

$$S_{G_1} = \left\{ \varphi \in \mathbb{R}^m | \beta = \begin{pmatrix} \varphi \\ 0 \end{pmatrix} \right\}. \tag{81}$$

The model with which we compare G_1, G_2, is identical to G, such that S_{G_2} reads

$$S_{G_2} = \{\beta \in \mathbb{R}^k\}. \tag{82}$$

Since $\sigma^2 \in \mathbb{R}^+$ is a nuisance parameter, we respecify G_1 and G_2 using the notation introduced in Section 7,

$$G_1: P(\sigma)y = P(\sigma)X\varphi + \varepsilon,$$

$$G_2: P(\sigma)y = P(\sigma)(X \;\; Z)\beta + \varepsilon, \tag{83}$$

where $P(\sigma) = \sigma^{-1} I_T$ and $\varepsilon \sim \mathrm{N}(0, I_T)$.

When we specify $g_i(\varphi_i, \lambda_i)$ as $g(\varphi, \lambda) = (0 \;\; I_{k-m})'\lambda$ with $\lambda: (k-m) \times 1$, Assumption 1 holds and the specification of β becomes

$$\beta = \begin{pmatrix} I_m \\ 0 \end{pmatrix} \varphi + \begin{pmatrix} 0 \\ I_{k-m} \end{pmatrix} \lambda. \tag{84}$$

Specification (84) satisfies Assumption 1 since it is an invertible relationship and (a) $\beta = (I_m \;\; 0)'\varphi \Leftrightarrow \lambda = 0$, (b) all values of φ lead to a unique value of β both when $\lambda = 0$ and when $\lambda \neq 0$, (c) $\partial g(\varphi, \lambda)/\partial \lambda' = (0 \;\; I_{k-m})'$ and does not depend on φ and λ. A prior specified on (β, σ^2) in model G therefore implies invariant prior/posterior probabilities for G_1 and G_2 and densities for φ when we apply Theorem 7.

Since $\partial\beta/\partial\varphi' \, \partial\lambda'$ does not depend on φ and λ, the priors and posteriors that result from Theorem 7 are identical to the priors and posteriors for φ that typically result when we specify them directly on φ. Hence, inducing the priors and posteriors on the parameters of linear models from the priors and posteriors on the parameters of ELR models does not lead to a notable difference for the resulting priors and posteriors compared to specifying them directly. This property of linear models has mistakenly been thought to hold for regression models that are non-linear in the parameters as well. However, for these models priors that are specified directly on the parameters do not correspond with the priors that are in general used on the parameters of an ELR model. Inducing the prior and posterior probability from the prior and posterior on the parameters of an ELR model does lead to a notable change of the prior and posterior probability in linear models.

8.2. Cointegration model

Cointegration implies a non-linear restriction on the parameters of a linear regression model. The restriction that cointegration implies is that the long run multiplier of a vector autoregressive model has a reduced rank value, see e.g. Engle and Granger (1987) and Johansen (1991). For a vector autoregressive model of order 1, cointegration with r cointegrating vectors implies that we can specify it as

$$G_r : \Delta y_t = \alpha_r' \beta_r' y_{t-1} + u_t, \quad r = 0, \ldots, k-1, \tag{85}$$

where y_t, $u_t : k \times 1$; $\Delta y_t = y_t - y_{t-1}$, α_r', $\beta_r : k \times r$, and u_t, $t = 1, \ldots, T$, are independently and identically normal distributed with mean zero and $k \times k$ covariance matrix Ω. In case r equals zero, $\alpha_0' \beta_0'$ is a $k \times k$ matrix of zeros. When we do not impose any normalization on α_r and β_r, the elements of α_r and β_r are non-identified since $\alpha_r' \beta_r' = \alpha_r^{*\prime} \beta_r^{*\prime}$, $\alpha_r^* = A\alpha_r$ and $\beta_r^* = \beta_r A^{-1}$ for any non-singular $r \times r$ matrix A. We therefore need to normalize either α_r and β_r, see Bauwens and Lubrano (1996). A straightforward normalization is to use $\beta_r = (I_r \;\; -\beta_{2,r}')'$ with $\beta_{2,r} : (k-r) \times r$. Because of the invariance of the Hausdorff-integrals to transformations, the chosen normalization has no consequences for the prior and posterior (probabilities) that result from the induced probability approach. Hence, the prior and posterior using one specific normalization are a transformation of the prior and posterior of another normalization, see Kleibergen and Paap (2002) for a proof. The prior and posterior probabilities are the same for all possible normalizations.

We represent the cointegration models G_r from (85) in matrix notation

$$G_r : Y = X\beta_r\alpha_r + U, \quad r = 0, \ldots, k-1, \tag{86}$$

where $Y = (\Delta y_1 \cdots \Delta y_T)'$, $X = (y_0 \cdots y_{T-1})'$, $U = (u_1 \cdots u_T)'$. The cointegration models G_r in (86) are nested in the multi-variate ELR model

$$G : Y = X\Pi + U \tag{87}$$

with $\Pi : k \times k$. We specify the cointegration models G_r in (86) and the ELR model G from (87) in line with Theorem 7 as

$$\begin{aligned} G_r &: P(\Omega)\,\mathrm{vec}(Y) = P(\Omega)(I_k \otimes X)\,\mathrm{vec}(\beta_r\alpha_r) + \mathrm{vec}(\varepsilon), \quad r = 0, \ldots, k-1, \\ G &: P(\Omega)\,\mathrm{vec}(Y) = P(\Omega)(I_k \otimes X)\,\mathrm{vec}(\Pi) + \mathrm{vec}(\varepsilon), \end{aligned} \tag{88}$$

where $P(\Omega)=(\Omega^{-1/2}\otimes I_T)$, $\varepsilon=U\Omega^{-1/2}$, $\text{vec}(\varepsilon)\sim \text{N}(0,I_{kT})$. Eq. (88) shows that G_r, $r=0,\ldots,k-1$, is represented by the lower-dimensional sets

$$S_{G_r}=\left\{\alpha_r\in\mathbb{R}^{k,r},\ \beta_{2,r}\in\mathbb{R}^{(k-r),r}|\Pi=\begin{pmatrix} I_r \\ -\beta_{2,r}\end{pmatrix}\alpha_r\right\},\qquad r=0,\ldots,k-1, \tag{89}$$

so S_{G_0} only consists of the $k\times k$ matrix of zeros.

The unrestricted full rank model G_k is identical to G such that S_{G_k} reads

$$S_{G_k}=\{\Pi\in\mathbb{R}^{k,k}\}. \tag{90}$$

The sets S_{G_r}, $r=0,\ldots,k$ are such that $S_{G_0}\subset S_{G_1}\subset\cdots\subset S_{G_k}$.

Since cointegration imposes a non-linear restriction on the parameters of a linear regression model, the specification of a function $g_i(\varphi_i,\lambda_i)$ that makes Assumption 1 hold is rather difficult to obtain. In Kleibergen and Paap (2002) a specification of Π that, results from a singular value decomposition and, satisfies Assumption 1 is given:

$$\Pi=\beta_r\alpha_r+\beta_{r,\perp}\lambda_r\alpha_{r,\perp},\qquad r=1,\ldots,k-1,$$

$$\Leftrightarrow$$

$$\text{vec}(\Pi)=\text{vec}(\beta_r\alpha_r)+(\alpha'_{r,\perp}\otimes\beta_{r,\perp})\,\text{vec}(\lambda_r),\qquad r=1,\ldots,k-1, \tag{91}$$

where $\lambda_r:(k-r)\times(k-r)$; $\beta_{r,\perp}$, $\alpha'_{r,\perp}:k\times(k-r)$ and $\beta'_{r,\perp}\beta_r\equiv 0$, $\beta'_{r,\perp}\beta_{r,\perp}\equiv I_{k-r}$, $\alpha_r\alpha'_{r,\perp}\equiv 0$, $\alpha_{r,\perp}\alpha'_{r,\perp}\equiv I_{k-r}$, such that

$$g_r(\varphi_r,\lambda_r)=(\alpha'_{r,\perp}\otimes\beta_{r,\perp})\,\text{vec}(\lambda_r),\qquad r=1,\ldots,k-1 \tag{92}$$

with $\varphi_r=(\alpha_r,\beta_{2,r})$. When r equals 0, $\lambda_0=\Pi$ since $\beta_0\alpha_0$ is a $k\times k$ matrix of zeros. Kleibergen and Paap (2002) show that an invertible relationship between Π and $(\alpha_r,\beta_{2,r},\lambda_r)$ exists. Furthermore, with respect to the conditions of Assumption 1: (a) $\Pi=\beta_r\alpha_r$ is equivalent to $\lambda_r=0$. (b) $(\alpha_r,\beta_{2,r})$ implies a unique value of $\beta_r\alpha_r$ when α_r has full rank and identically $(\alpha_r,\beta_{2,r},\lambda_r)$ implies a unique value of $\beta_r\alpha_r+\beta_{r,\perp}\lambda_r\alpha_{r,\perp}$ when α_r has full rank. (c) $\partial g_r(\varphi_r,\lambda_r)/\partial\,\text{vec}(\lambda_r)'=(\alpha'_{r,\perp}\otimes\beta_{r,\perp})$ such that $(\partial g_r(\varphi_r,\lambda_r)/\partial\,\text{vec}(\lambda_r)')'(\partial g_r(\varphi_r,\lambda_r)/\partial\,\text{vec}(\lambda_r)')=I_{(k-r)^2}$ and does not depend on φ_r and λ_r. Hence, all conditions of Assumption 1 are satisfied. Theorem 7 therefore applies and a prior specified on (Π,Ω) in G implies a prior probability for G_r, $r=0,\ldots,k$, and a prior for $(\alpha_r,\beta_{2,r},\Omega)$ in G_r that are invariant with respect to the specification of Π and $(\alpha_r,\beta_{2,r},\lambda_r)$ that satisfy Assumption 1. For more details about the resulting Bayesian analysis of the cointegration model and the expressions of the priors and posteriors, we refer to Kleibergen and Paap (2002).

8.3. ARMA(1,1)

Another model that results from a non-linear restriction on the parameters of an ELR model is the ARMA(1,1) model, see e.g. Box et al. (1994),

$$G_1: y_t=\rho y_{t-1}-\alpha u_{t-1}+u_t,\qquad t=1,\ldots,T, \tag{93}$$

where the disturbances u_t are independently and identically distributed, $u_t \sim \mathrm{N}(0, \sigma^2)$. When we recurrently substitute u_{t-1} in (93), we obtain

$$G_1 : y_t = (\rho - \alpha) \sum_{j=1}^{T} \alpha^{j-1} y_{t-j} + u_t, \quad t = 1, \ldots, T. \tag{94}$$

We specify (94) as a regression model that is non-linear in the parameters,

$$G_1 : y = Xf(\alpha, \rho) + u, \tag{95}$$

where $y = (y_1 \cdots y_T)'$, $X = (x_1 \cdots x_T)'$, $x_i = (y_{i-1} \cdots y_0 \; 0 \cdots 0)' : T \times 1$, $i = 1, \ldots, T$; $u = (u_1 \cdots u_T)'$, and

$$f(\alpha, \rho) = (\rho - \alpha) \begin{pmatrix} 1 \\ \alpha \\ \vdots \\ \alpha^{T-1} \end{pmatrix} : T \times 1. \tag{96}$$

G_1 in (95) is nested in the ELR model

$$G : y = X\beta + u \tag{97}$$

with $\beta : T \times 1$. We specify both G_1 and G in the notation of Theorem 7,

$$G_1 : P(\sigma)y = P(\sigma)Xf(\alpha, \rho) + \varepsilon,$$
$$G : P(\sigma)y = P(\sigma)X\beta + \varepsilon \tag{98}$$

with $P(\sigma) = \sigma^{-1} I_T$ and $\varepsilon \sim \mathrm{N}(0, I_T)$.

The ARMA(1,1) model imposes a non-linear restriction on the parameters of G, $\beta = f(\alpha, \rho)$. This implies that it is not straightforward to obtain a specification of $g_i(\varphi_i, \lambda_i)$ that satisfies Assumption 1. A (unrestricted) specification of β that gives such a function $g_i(\varphi_i, \lambda_i)$ is

$$\beta = (\rho - \alpha) \begin{pmatrix} 1 \\ \alpha \\ \vdots \\ \alpha^{T-1} \end{pmatrix} + \begin{pmatrix} 0 \\ I_{T-2} \end{pmatrix} \lambda \tag{99}$$

with $\lambda : (T-2) \times 1$ and $g(\varphi, \lambda) = (0 \; I_{T-2})'\lambda$ with $\varphi = (\alpha, \rho)$. Eq. (99) satisfies the conditions from Assumption 1 since: β has an invertible relationship with (α, ρ, λ), (a) $\beta = f(\alpha, \rho) \Leftrightarrow \lambda = 0$, (b) (ρ, α) implies a unique value of $f(\alpha, \rho)$ when $\rho - \alpha \neq 0$, identically (ρ, α, λ) implies a unique value of $f(\alpha, \rho) + (0 \; I_{T-2})'\lambda$ when $\rho - \alpha \neq 0$, (c) $\partial g(\varphi, \lambda)/\partial \lambda' = (0 \; I_{T-2})'$ and independent of (α, ρ, λ). Theorem 7 therefore applies and a prior that is specified on (β, σ^2) in G induces a prior probability for G_1 and a prior on

(α, ρ, σ^2) that are invariant with respect to the specification of β and (α, ρ, λ) that satisfy Assumption 1. For more details on the resulting Bayesian analysis of the ARMA(1,1) model, we refer to Kleibergen and Hoek (1999).

9. Conclusions

We obtain expressions for prior/posterior probabilities and densities on the parameters of nested regression models that are induced by the prior/posterior on the parameters of an encompassing linear regression model. The resulting probabilities and densities are invariant with respect to specifications that satisfy a necessary set of assumptions. Hence, by specifying a prior and a likelihood for the parameters of an encompassing linear regression model, we obtain a complete Bayesian analysis, that includes both prior/posterior probabilities and densities, for all of its nested regression models that allow for a specification that satisfies the set of assumptions. The resulting Bayesian analyzes of these nested regression models are consistent with one another.

The Bayes factor in the resulting analysis corresponds with the Savage–Dickey density ratio and equals the ratio of the posterior and prior height in the hypothesized parameter point. When we multiply the Bayes factor with the prior odds ratio, we obtain the posterior odds ratio. Because the prior and prior probability result from the same prior on the parameters of the encompassing linear regression model, the posterior odds ratio is such that the prior odds ratio corrects the Bayes factor for the plausibility of the competing models reflected in the prior. The posterior odds ratio is therefore robust to increases in the prior variance which is an element of the Jeffreys–Lindley's paradox.

Applications of the above results are especially important for regression models that result from non-linear restrictions on the parameters of encompassing linear regression models. In these models, the resulting analysis leads to priors and posteriors that are different from the ones that are used traditionally. The traditional Bayesian analysis leads to anomalies in these models, like, for example, in simultaneous equation models, see Kleibergen (1997) and Kleibergen and van Dijk (1998); cointegration models, see Kleibergen and van Dijk (1994) and Martin and Martin (2000); and fractional cointegration models, see Martin (2001). When we deduce the priors and posteriors of the parameters in these models from priors and posteriors on the parameters of encompassing linear regression models, these anomalies disappear, see e.g. Kleibergen (1997), Kleibergen and van Dijk (1998) and Kleibergen and Paap (2002).

Acknowledgements

I thank the editors, participants of the conference on "Recent advances in Bayesian Econometrics" in June 2001 at GREQAM Marseille and especially Jacek Osiewalski and three anonymous referees for helpful comments and suggestions. This research has partly been funded by the NWO Vernieuwingsimpuls research grant "Empirical Comparison of Economic Models".

Appendix A.

Proof of Eq. (11). The normalizing constant is specified as the transformation over g_i of the $(k-m_i)$-dimensional sphere with center zero and radius ρ:

$$\begin{aligned} c_i(\rho)^{-1} &= L_{k-m_i}(g_i(0, B_i(0,\rho))) \\ &= \int_{B_{k-m_i}(0,\rho)} \left|\frac{\partial g_i}{\partial \lambda_i}\right| L_{k-m_i}(\mathrm{d}\lambda_i) \\ &= \int_{B_{k-m_i}(0,\rho)} \left|\left(\frac{\partial g_i}{\partial \lambda_i}\right)'\left(\frac{\partial g_i}{\partial \lambda_i}\right)\right|^{1/2} \mathrm{d}\lambda_i \\ &= \int_{B_{k-m_i}(0,\rho)} |A_i|^{1/2}\,\mathrm{d}\lambda_i \\ &= |A_i|^{1/2}\int_{B_{k-m_i}(0,\rho)} \mathrm{d}\lambda_i \\ &= |A_i|^{1/2} V_{k-m_i}(\rho) \end{aligned}$$

with $V_{k-m_i}(\rho)$ the volume of the $(k-m_i)$-dimensional sphere with radius ρ. Because $|A_i|$ does not depend on φ_i and λ_i, the value of φ_i in $g_i(\varphi_i, \lambda_i)$ and the center of the sphere $B_i(0,\rho)$ do not affect the expression of $c_i(\rho)$ and we have therefore used vectors of zeros for convenience. □

Proof of Theorem 1. Before we obtain the specification of the Hausdorff-measure, we note the structure that Assumption 1 imposes on the Jacobian of the transformation from β to (φ_i, λ_i):

$$J(\beta,(\varphi_i,\lambda_i)) = \left(\frac{\partial f_i}{\partial \varphi_i'} + \frac{\partial g_i}{\partial \varphi_i'}\ \ \frac{\partial g_i}{\partial \lambda_i'}\right).$$

Because $g_i(\varphi_i,\lambda_i)$ is a strictly monotonic function of λ_i and $g_i(\varphi_i,\lambda_i)=0 \Leftrightarrow \lambda_i = 0$, $\partial g_i/\partial \varphi_i'|_{\lambda_i=0} = 0$. Hence, the Jacobian in $\lambda_i = 0$ reads

$$J(\beta,(\varphi_i,\lambda_i))|_{\lambda_i=0} = \left(\frac{\partial f_i}{\partial \varphi_i'}\ \ \left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)$$

and

$$\begin{aligned} |J(\beta,(\varphi_i,\lambda_i))|_{\lambda_i=0}| &= \left|\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)'\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)\right|^{1/2} \left|\left(\frac{\partial f_i}{\partial \varphi_i'}\right)' M_{\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)}\left(\frac{\partial f_i}{\partial \varphi_i'}\right)\right|^{1/2} \\ &= \left|\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)' M_{\left(\frac{\partial f_i}{\partial \varphi_i'}\right)}\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)\right|^{1/2} \left|\left(\frac{\partial f_i}{\partial \varphi_i'}\right)'\left(\frac{\partial f_i}{\partial \varphi_i'}\right)\right|^{1/2}, \end{aligned}$$

where

$$\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)'\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)=\left(\frac{\partial g_i}{\partial \lambda_i'}\right)'\left(\frac{\partial g_i}{\partial \lambda_i'}\right)=A_i.$$

The Hausdorff-measure $H_{m_i}(W_{G_i})$ is obtained by considering that $g_i(\varphi_i,\lambda_i)$ is a strictly monotonic function of λ_i. We use the sequence of sets $W_{G_i}(\rho)$ centered at $\lambda_i=0$,

$$W_{G_i}(\rho)=\{\varphi_i\in\Omega_{G_i}\subset\mathbb{R}^{m_i},\ \lambda_i\in B_{k-m_i}(0,\rho)\subset\mathbb{R}^{k-m_i}|\beta=f(\varphi_i)+g_i(\varphi_i,\lambda_i)\},$$

where $B_{k-m_i}(0,\rho)$ is a $(k-m_i)$-dimensional sphere with radius ρ centered at 0. We use a limiting sequence of $W_{G_i}(\rho)$ that is obtained by letting ρ converge to zero,

$$\lim_{\rho\to 0} W_{G_i}(\rho)=W_{G_i}.$$

This results because $g_i(\varphi_i,\lambda_i)$ is a strictly monotonic function of λ.

Because

$$\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)'\left(\left.\frac{\partial g_i}{\partial \lambda_i'}\right|_{\lambda_i=0}\right)=A_i \quad\text{and}\quad \left.\frac{\partial g_i}{\partial \varphi_i'}\right|_{\lambda_i=0}=0,$$

the Lebesgue-measure of $W_{G_i}(\rho)$, $L_k(W_{G_i}(\rho))$, is for small values of ρ equal to

$$\begin{aligned}L_k(W_{G_i}(\rho))&=\int_{\Omega_{G_i}}\int_{B_{G_i}(0,\rho)}|J(\beta,(\varphi_i,\lambda_i))|\,\mathrm{d}\lambda_i\,\mathrm{d}\varphi_i\\
&\approx\int_{\Omega_{G_i}}\left|\left(\frac{\partial f_i}{\partial\varphi_i'}\right)' M_{\left(\left.\frac{\partial g_i}{\partial\lambda_i'}\right|_{\lambda_i=0}\right)}\left(\frac{\partial f_i}{\partial\varphi_i'}\right)\right|^{1/2}\left[\int_{B_{k-m_i}(0,\rho)}|A_i|^{1/2}\,\mathrm{d}\lambda_i\right]\mathrm{d}\varphi_i\\
&\approx\left\{\int_{\Omega_{G_i}}\left|\left(\frac{\partial f_i}{\partial\varphi_i'}\right)' M_{\left(\left.\frac{\partial g_i}{\partial\lambda_i'}\right|_{\lambda_i=0}\right)}\left(\frac{\partial f_i}{\partial\varphi_i'}\right)\right|^{1/2}\mathrm{d}\varphi_i\right\}|A_i|^{1/2}V_{k-m_i}(\rho).\end{aligned}$$

The Hausdorff-measure equals the limit of $c_i(\rho)$ times $L_k(W_{G_i}(\rho))$ when ρ converges to zero:

$$\begin{aligned}H_{m_i}(W_{G_i})&=\lim_{\rho\to 0}[c_i(\rho)L_k(W_{G_i}(\rho))]\\
&=\lim_{\rho\to 0}\frac{1}{|A_i|^{1/2}V_{k-m_i}(\rho)}\left\{\int_{\Omega_{G_i}}\left|\left(\frac{\partial f_i}{\partial\varphi_i'}\right)' M_{\left(\left.\frac{\partial g_i}{\partial\lambda_i'}\right|_{\lambda_i=0}\right)}\left(\frac{\partial f_i}{\partial\varphi_i'}\right)\right|^{1/2}\mathrm{d}\varphi_i\right\}\\
&\quad\times|A_i|^{1/2}V_{k-m_i}(\rho)\\
&=\int_{\Omega_{G_i}}\left|\left(\frac{\partial f_i}{\partial\varphi_i'}\right)' M_{\left(\left.\frac{\partial g_i}{\partial\lambda_i'}\right|_{\lambda_i=0}\right)}\left(\frac{\partial f_i}{\partial\varphi_i'}\right)\right|^{1/2}\mathrm{d}\varphi_i.\end{aligned}$$

To show the invariance of the Hausdorff-measure, we consider an invertible function $h:\mathbb{R}^k \to \mathbb{R}^k$, $\mu = h(\beta)$. Because of Assumption 1, we can specify β as

$$\beta = f_i(\varphi_i) + g_i(\varphi_i, \lambda_i)$$

and μ can therefore be specified as

$$\mu = l_i(\psi_i) + r_i(\psi_i, \theta_i)$$

with $l_i(\psi_i) = h(f_i(\varphi_i))$ and $r(\psi_i, \theta_i) = h(f_i(\varphi_i) + g_i(\varphi_i, \lambda_i)) - h(f_i(\varphi_i))$. Because of Assumption 1(b), that $g_i(\varphi_i, \lambda_i)$ is a strictly monotonic function of λ_i, h has to be strict monotonic. This implies that $(\partial h/\partial\beta')'(\partial h/\partial\beta')$ is a positive definite symmetric matrix for all values of β and that θ_i is an invertible function of λ_i only. Because of Assumption 1(c),

$$\frac{\partial r_i}{\partial \theta_i'} = \frac{\partial \mu}{\partial \beta'}\frac{\partial \beta}{\partial \lambda'}\frac{\partial \lambda}{\partial \theta_i'} = \frac{\partial h}{\partial \beta'}\frac{\partial g_i}{\partial \lambda_i'}\frac{\partial \lambda_i}{\partial \theta_i'}$$

should be such that

$$\left(\frac{\partial r_i}{\partial \theta_i'}\right)'\left(\frac{\partial r_i}{\partial \theta_i'}\right) = B_i \quad \Leftrightarrow$$

$$\left(\frac{\partial h}{\partial \beta'}\left(\frac{\partial g_i}{\partial \lambda_i'}\right)\left(\frac{\partial \lambda_i}{\partial \theta_i'}\right)\right)'\left(\frac{\partial h}{\partial \beta'}\left(\frac{\partial g_i}{\partial \lambda_i'}\right)\left(\frac{\partial \lambda_i}{\partial \theta_i'}\right)\right) = B_i \quad \Leftrightarrow$$

$$\left(\frac{\partial \lambda_i}{\partial \theta_i'}\right)'\left(\frac{\partial g_i}{\partial \lambda_i'}\right)'\left(\frac{\partial h}{\partial \beta'}\right)'\left(\frac{\partial h}{\partial \beta'}\right)\left(\frac{\partial g_i}{\partial \lambda_i'}\right)\left(\frac{\partial \lambda_i}{\partial \theta_i'}\right) = B_i$$

with B_i independent of ψ_i. Since θ_i is an invertible function of λ_i only and $(\partial g_i/\partial\lambda_i')'(\partial g_i/\partial\lambda_i') = A_i$, with A_i independent of (φ_i, λ_i), $(\partial h/\partial\beta')'(\partial h/\partial\beta')$ should be equal to some fixed positive definite symmetric matrix that is independent of β. Unlike g_i, h is an invertible function such that the only specification of h that satisfies all conditions is an invertible linear function. Hence, every specification $\mu = l_i(\psi_i) + r_i(\psi_i, \theta_i)$ that satisfies Assumption 1 is such that (1) μ is an invertible linear function of β and (2) θ_i is an invertible function of λ_i only and ψ_i is an invertible function of φ_i only. It is straightforward to show that these transformations lead to an identical Hausdorff-measure. □

Proof of Theorem 2. For small values of ρ, the expression of $\int_{W_{G_i}(\rho)} q(\beta)\,\mathrm{d}\beta$ reads

$$\begin{aligned}\int_{W_{G_i}(\rho)} q(\beta)\,\mathrm{d}\beta &= \int_{\Omega_{G_i}}\int_{B_{k-m_i}(0,\rho)} q(\beta(\varphi_i,\lambda_i))|J(\beta,(\varphi_i,\lambda_i))|\,\mathrm{d}\lambda_i\,\mathrm{d}\varphi_i \\ &\approx \int_{\Omega_{G_i}}\left[\int_{B_{k-m_i}(0,\rho)} q(\beta(\varphi_i,\lambda_i)|_{\lambda_i=0})|J(\beta,(\varphi_i,\lambda_i))|_{\lambda_i=0}|\,\mathrm{d}\lambda_i\right]\mathrm{d}\varphi_i \\ &\approx \left\{\int_{\Omega_{G_i}} q(\beta(\varphi_i,\lambda_i)|_{\lambda_i=0})|J(\beta,(\varphi_i,\lambda_i))|_{\lambda_i=0}|\,\mathrm{d}\varphi_i\right\} V_{k-m_i}(\rho).\end{aligned}$$

The Hausdorff-integral is then obtained by

$$\int_{W_{G_i}} q(\beta)H_{m_i}(\mathrm{d}\beta) = \lim_{\rho\to 0}\left[c_i(\rho)\int_{W_{G_i}(\rho)} q(\beta)\,\mathrm{d}\beta\right]$$
$$= \lim_{\rho\to 0}\left[\frac{\{\int_{\Theta_{G_i}} q(\beta(\varphi_i,\lambda_i)|_{\lambda_i=0})|J(\beta,(\varphi_i,\lambda_i))|_{\lambda_i=0}|\,\mathrm{d}\varphi_i\}V_{k-m_i}(\rho)}{|A_i|^{1/2}V_{k-m_i}(\rho)}\right]$$
$$= \frac{1}{|A_i|^{1/2}}\left\{\int_{\Theta_{G_i}} q(\beta(\varphi_i,\lambda_i)|_{\lambda_i=0})|J(\beta,(\varphi_i,\lambda_i))|_{\lambda_i=0}|\,\mathrm{d}\varphi_i\right\}.$$

The proof of the invariance of the Hausdorff-integral to specifications of β, (φ_i,λ_i) that satisfy Assumption 1 is analogous to the proof for Theorem 1. □

Proof of Theorem 4. Eq. (28) gives the definition of a density function. The invariance of it follows from the proof of Theorem 1. We have shown in this proof that when

$$\beta = f_i(\varphi_i) + g_i(\varphi_i,\lambda_i)$$

and

$$\mu = l_i(\psi_i) + r_i(\psi_i,\theta_i)$$

are two specifications that satisfy Assumption 1 that ψ_i is an invertible function of φ_i only and θ_i is an invertible function of λ_i only. Hence, we can independently transform φ_i to ψ_i and λ_i to θ_i. This does not affect the specification of the prior from Theorem 4. □

Proof of Eq. (44).

$$p_{G_i}(D) = \int_{\Theta_{G_i}} p_{G_i}(\varphi_i)\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}\,\mathrm{d}\varphi_i$$
$$= p_G(\lambda_i|D)|_{\lambda_i=0}\left\{\int_{\Theta_{G_i}}\left[\frac{p_{G_i}(\varphi_i)\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}}{p_G(\lambda_i|D)|_{\lambda_i=0}}\right]\mathrm{d}\varphi_i\right\}$$
$$= p_G(\lambda_i|D)|_{\lambda_i=0}\left\{\int_{\Theta_{G_i}}\left[\frac{p_{G_i}(\varphi_i)\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}p_G(\varphi_i|\lambda_i,D)|_{\lambda_i=0}}{p_G(\lambda_i|D)|_{\lambda_i=0}\,p_G(\varphi_i|\lambda_i,D)|_{\lambda_i=0}}\right]\mathrm{d}\varphi_i\right\}$$
$$= p_G(\lambda_i|D)|_{\lambda_i=0}\left\{\int_{\Theta_{G_i}}\left[\frac{p_{G_i}(\varphi_i)\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}p_G(\varphi_i|\lambda_i,D)|_{\lambda_i=0}}{p_G(\varphi_i,\lambda_i|D)|_{\lambda_i=0}}\right]\mathrm{d}\varphi_i\right\}$$
$$= p_G(\lambda_i|D)|_{\lambda_i=0}\left\{\int_{\Theta_{G_i}}\left[\frac{p_{G_i}(\varphi_i)\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}p_G(\varphi_i|\lambda_i,D)|_{\lambda_i=0}}{\frac{p_G(\varphi_i,\lambda_i)|_{\lambda_i=0}\mathscr{L}(D|\beta)|_{\beta=f_i(\varphi_i)}}{c_\beta}}\right]\mathrm{d}\varphi_i\right\}$$
$$= c_\beta\times p_G(\lambda_i|D)|_{\lambda_i=0}\left\{\int_{\Theta_{G_i}}\left[\frac{p_{G_i}(\varphi_i)p_G(\varphi_i|\lambda_i,D)|_{\lambda_i=0}}{p_G(\varphi_i,\lambda_i)|_{\lambda_i=0}}\right]\mathrm{d}\varphi_i\right\}$$

$$= c_\beta \times \frac{p_G(\lambda_i|D)|_{\lambda_i=0}}{\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}\,\mathrm{d}\varphi_i}$$

$$\times \left\{ \int_{\Theta_{G_i}} \left[\frac{p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}\, p_G(\varphi_i|\lambda_i, D)|_{\lambda_i=0}}{p_G(\lambda_i)|_{\lambda_i=0}\, p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}} \right] \mathrm{d}\varphi_i \right\}$$

$$= c_\beta \times \frac{p_G(\lambda_i|D)|_{\lambda_i=0}}{p_G(\lambda_i)|_{\lambda_i=0}} \times \frac{\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i, D)|_{\lambda_i=0}\,\mathrm{d}\varphi_i}{\int_{\Theta_{G_i}} p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}\,\mathrm{d}\varphi_i},$$

where

$$c_\beta^{-1} = \int_{\mathbb{R}^k} p_G(\beta)\mathscr{L}(D|\beta)\,\mathrm{d}\beta$$

$$= \int_{\mathbb{R}^{m_i}} \int_{\mathbb{R}^{k-m_i}} p_G(\varphi_i, \lambda_i)\mathscr{L}(D|\beta(\lambda_i, \varphi_i))\,\mathrm{d}\lambda_i\,\mathrm{d}\varphi_i,$$

$$p_{G_i}(\varphi_i) = \frac{p_G(\varphi_i|\lambda_i)|_{\lambda_i=0}}{\int_{\Theta_{G_i}} p_G(u|\lambda_i)|_{\lambda_i=0}\,\mathrm{d}u}. \qquad \square$$

References

Bauwens, L., Lubrano, M., 1996. Identification restrictions and posterior densities in cointegrating Gaussian VAR systems. In: Fomby, T. (Ed.), Advances in Econometrics. JAI Press, Greenwich, CT.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Inference. Springer, New York.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley, New York.

Billingsley, P., 1986. Probability and Measure. Wiley, New York.

Box, G.E.P., Tiao, G.C., 1992. Bayesian Inference in Statistical Analysis. Wiley, Wiley classics library edition edn.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis, 3rd Edition. Prentice-Hall, New Jersey.

Chao, J.C., Phillips, P.C.B., 1998. Posterior distributions in limited information analysis of the simultaneous equations models using the Jeffreys' prior. Journal of Econometrics 87, 49–86.

De Finetti, B., 1972. Probability, Induction and Statistics. Wiley, New York.

Dickey, J.M., 1971. The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. The Annals of Mathematical Statistics 42, 204–223.

Doster, W., 1998. Jeffrey's prior is the Hausdorff measure for the Hellinger and Kullback-Leibler distances. Working paper, Department of Mathematics, University of Kaiserslautern.

Drèze, J.H., Richard, J.F., 1983. Bayesian analysis of simultaneous equations systems. In: Griliches, Z., Intrilligator, M.D. (Eds.), Handbook of Econometrics, Vol. 1. Elsevier Science, Amsterdam.

Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation and testing. Econometrica 55, 251–276.

Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59, 1551–1580.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kleibergen, F., 1997. Bayesian simultaneous equations analysis using equality restricted random variables. In: 1997 Proceedings of the Section on Bayesian Statistical Science, American Statistical Association, pp. 141–146.

Kleibergen, F., Hoek, H., 1999. Bayesian Analysis of ARMA models. Tinbergen Institute Discussion Paper TI 2000-027/4.

Kleibergen, F., Paap, R., 2002. Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration. Journal of Econometrics 111, 223–249.
Kleibergen, F., van Dijk, H.K., 1994. On the shape of the likelihood/posterior in cointegration models. Econometric Theory 10, 514–551.
Kleibergen, F., van Dijk, H.K., 1998. Bayesian simultaneous equation analysis using reduced rank structures. Econometric Theory 14, 701–743.
Kleibergen, F., Zivot, E., 2003. Bayesian and classical approaches to instrumental variable regression. Journal of Econometrics 114, 29–72.
Kolmogorov, A.N., 1950. Foundations of the Theory of Probability. Chelsea, New York.
Lindley, D.V., 1957. A statistical paradox. Biometrika 44, 187–192.
Martin, G.M., 2001. Bayesian analysis of a fractional cointegration model. Econometric Reviews 20, 217–234.
Martin, G.M., Martin, V.L., 2000. Bayesian inference in the triangular cointegration model using a Jeffreys prior. Communications in Statistics, Theory and Methods 29 (8), 1759–1785.
McCulloch, R.E., Rossi, P.E., 1992. Bayes factors for nonlinear hypotheses and likelihood distributions. Biometrika 79, 663–676.
O'Hagan, A., 1994. Bayesian Theory. In: Kendall's Advanced Theory of Statistics, Vol. 2B. Arnold, Paris.
Phillips, P.C.B., 1996. Econometric model determination. Econometrica 64, 763–812.
Phillips, P.C.B., Ploberger, W., 1994. Posterior odds testing for a unit root with data-based model selection. Econometric Theory 10, 774–808.
Phillips, P.C.B., Ploberger, W., 1996. An asymptotic theory of Bayesian inference for time series. Econometrica 64, 381–412.
Poirier, D.J., 1995. Intermediate Statistics and Econometrics: A Comparative Approach. MIT Press, Cambridge, MA.
Rogers, C.A., 1999. Hausdorff measures, 2nd Edition. Cambridge University Press, Cambridge, MA.
Tiao, G.C., Tan, W.-Y., Chang, Y.-C., 1977. Some aspects of bivariate regression subject to linear constraints. Journal of Econometrics 5, 13–35.
Verdinelli, I., Wasserman, L., 1995. Computing Bayes factors using a generalization of the Savage–Dickey density ratio. Journal of the American Statistical Association 90, 614–618.
Wolpert, R.J., 1995. Comment on: inference for a deterministic population dynamics model for Bowhead Whales by, A.E. Raftery, G.H. Rivens and J.E. Zeh. Journal of the American Statistical Association 90, 426–427.