A Bayesian Perspective on Model Selection

Jacek Osiewalski
Academy of Economics
Kraków, Poland

Mark F.J. Steel
Tilburg University
Tilburg, The Netherlands

original English version
of the paper

Una perspectiva bayesiana en selección de modelos

published in

Cuadernos Economicos 55/3 (1993), pp. 327-351

A Bayesian Perspective on Model Selection

Jacek Osiewalski
Academy of Economics
Kraków, Poland

Mark F.J. Steel
Tilburg University
Tilburg, The Netherlands

1 Introduction

The problem of model selection in regression has been the focus of quite a number of recent Bayesian papers, such as e.g. Consonni and Veronese (1992), Gaver and Geisel (1974, 1976), Geweke (1988), Lempers (1971), Pericchi (1984), Pettit (1992), Poirier (1985, 1988a), Zellner and Siow (1980), and Zellner (1984).

Rather than attempting to give an exhaustive survey of the Bayesian literature in choosing between contending regression models, we shall briefly explain some guiding principles of Bayesian testing and examine their implications for model selection. Our main aim is to provide the applies researcher with a basic understanding of these procedures and to indicate how they can be used in actual econometric practice. We focus our presentation upon applied non-Bayesian readers. Therefore, we avoid technical details and provide a brief introduction to the Bayesian paradigm in Section 2. However, we hope that there is still sufficient material of interest to Bayesian readers, who might find it interesting to see these matters surveyed and put in perspective. Section 3 introduces three Bayesian testing principles: posterior odds, Lindley type testing and model encompassing. We discuss their motivation and applicability in the context of having to choose one model as our "favourite" (pretesting). The posterior odds principle, however, naturally leads to an approach where inferences on common parameters and observables are not made on the basis of one particular model, but through a weighted average of all contending models. This pooling approach is discussed in Section 4. Section 5 deals with the, so-called, choice of regressors problem, where the regression models considered differ in the means but the covariance structure is given. This has been the focus of most of the references cited above. However, in applied econometrics we often face the opposite situation where the mean is agreed upon, but the covariance structure is not. Examples are cases where one might expect serial correlation or heteroskedasticity. In the literature this was examined in Lempers (1971, Ch. 4), Poiries (1988a) and Osiewalski and Steel (1992, 1993a) and here Section 6 is devoted to this problem. We use posterior odds and an approximate Lindley type approach within a dynamic framework and apply this in Section 7 to a study of the influence of advertising on sales. Concluding remarks are grouped in Section 8.

2 Bayesian models and inference rules

The concept of a sampling model, i.e. a family of probability distributions of the observables $\tilde{y} \in \tilde{Y}$ indexed by some parameter $\omega \in \Omega$, is the common starting point of both the sampling-theory and Bayesian approaches to statistical modelling.[1] If, additionally, our lack of knowledge or uncertainty about ω is formalized through a distribution on the parameter space Ω, we are entering the Bayesian world.

Thus, the Bayesian model is defined as a joint distribution on the product of the sample space $\tilde{Y}$ and the parameter space Ω, which, in the case of existence of densities, can be represented as

$$p(\tilde{y}, \omega) = p(\tilde{y}|\omega)p(\omega), \tag{2.1}$$

where in $p(\tilde{y}, \omega)$ we recognize the sampling density and $p(\omega)$ is the prior density. This prior density captures the opinions on ω existing before having observed the data. From the moment that the Bayesian model has been formulated, all inference questions can be answered by the application of basic rules of probability calculus. In (2.1) we can condition on the observed data and we can marginalize with respect to all the remaining quantities expect the quantity of interest.

In order to cover prediction as well as parameter estimation, assume that $\tilde{y} = (y' \, y_f')'$, where $y \in Y$ is observed, $y_f \in Y_f$ is unobserved (to be forecasted), and $\tilde{Y} = Y \times Y_f$. Bayesian inference is based on different factorizations of the joint density (2.1), namely

$$\begin{aligned} p(y, y_f, \omega) &= p(y_f|y, \omega)p(y|\omega)p(\omega) \\ &= p(y_f|y, \omega)p(\omega|y)p(y) \\ &= p(y_f|y)p(y)p(\omega|y_f, y). \end{aligned} \tag{2.2}$$

[1] We assume that ω is a finitely dimensional vector, i.e. we restrict attention to parametric models only.

Parameter inference is based in the posterior density function $p(\omega|y) =$ $p(y|\omega)p(\omega)\ p(y)$, while prediction is based on the out-of-sample predictive density function $p(y_f|y)$. Our predictive inference then fully reflects our inherent uncertainty regarding ω, given the choice of a sampling model (and a prior density). The uncertainty about ω is formalized through the posterior density, which is clearly seen if we write $p(y_f|y) = \int_\Omega p(y_f|y,\omega)p(\omega|y)d\omega$. Both posterior and out-of-sample predictive distributions condition on the observed data. If a particular function of ω, say φ, is the parameter of interest, then its posterior density function, $p(\varphi|y)$, is derived from $p(\omega|y)$. Similarly, if one particular element of y_f is of interest, its marginal density is obtained from $p(y_f|y)$.

Whereas the frequentist or sampling-theory paradigm is based on the sampling properties in $\tilde{Y}$ given unknown, but fixed, parameter values ω, a Bayesian considers the probability distribution of ω and y_f given the observed values of y without taking into account what could have been observed in repeated sampling.

On the formal level, introducing a distribution over the parameter space and conditioning on the actual observations are the distinctive features of the Bayesian approach to inference. Also, the subjectivist interpretation of probability as a measure of degree of belief (or uncertainty) is widely adopted by Bayesian statisticians and econometriccians. Thus, any conclusion from Bayesian inference can be formulated in an intuitively straightforward manner, e.g. “given the data and the prior information, one can be 90% sure that φ is greater than φ_0”. For a more in-depth treatment of the differences between the classical and the Bayesian paradigm, see, e.g., Poirier (1988b).

3 Model choice procedures

Let us consider m competing sampling models defined on the same space $\tilde{Y}$,

$$M_i : p_i(\tilde{y}|\omega_i) = p_i(\tilde{y}|\theta,\eta_i), i = 1,\dots,m, \tag{3.1}$$

where $\omega_i = (\theta'\eta_i')' \in \Omega_i = \Theta \times H_i$ denotes all the parameters of M_i, while θ groups the parameters common to all m models and the η_i’s denote model specific parameters. Defining m prior distributions $p_i(\omega_i) = p_i(\theta,\eta_i)$, we obtain m Bayesian models:

$$p_i(\tilde{y},\omega_i) = p_i(\tilde{y}|\omega_i)p_i(\omega_i)$$

$$= p_i(y_f|y,\omega_i)p_i(y|\omega_i)p_i(\omega_i), i = 1, \dots, m, \tag{3.2}$$

where the individual model-specific posterior and predictive inference can be conducted following the general rules described in the previous section.

The formal probabilistic approach to model choice is based on posterior model probabilities. This leads us directly to the so-called posterior odds approach.

Assume that $M_1, \dots, M_m$ are mutually exclusive (non-nested) and jointly exhaustive. The reason for assuming a non-nested structure of the models is not inspired by any requirement of posterior odds testing *per se*. However, as we are going to attach prior probabilities $p(M_i)$ to all models, we logically need that $p(M_i) \geq p(M_j)$ if M_i nests M_j. This complicates the elicitation of prior model probabilities and often leads to discontinuities in the prior $p(\eta_i|\theta)$. If we assign prior probability $p(M_i)$ to the i-th model, then the posterior probability of M_i is given by

$$p(M_i|y) = \frac{p(M_i)p(y|M_i)}{\sum_{j=1}^{m} p(M_j)p(y|M_j)}, \tag{3.3}$$

where

$$p(y|M_i) = p_i(y) = \int_{\Omega_i} p_i(y|\omega_i)p_i(\omega_i)d\omega_i \, , i = 1, \dots, m, \tag{3.4}$$

is the within-sample predictive density (or the marginal data density) under M_i.

The Bayesian counterpart of the conventional pretest procedure is to first select a particular model by employing (3.3) and then conduct inference with the chosen model. A Bayesian approach fits directly into a formal decision theory framework. Generally, if we can take actions denoted by $a \in A$, then the consequences of such actions typically depend on parameters, e.g., $\omega \in \Omega$ and possibly (also) on future $y_f \in Y_f$. This can be formalized in a loss function $L(\omega, y_f, a)$ so that the posterior expected loss is easily calculated as

$$\int_{\Omega \times Y_f} L(\omega, y_f, a) \, p(\omega, y_f|y) \, d\omega \, dy_f \text{ for every } a \in A.$$

See Kiefer and Richard (1987) for a clear exposition in the case $L(\omega, y_f, a) = L(\omega, a)$ and Berger (1985) for a very complete treatment of decision theory. In the pretest approach the model choice that minimizes this posterior expected loss is suggested. If losses of incorrect decisions are identical then this is equivalent to the criterion of highest posterior model probability. Alternative loss structures, leading to different decisions, can be adopted, as in Monahan (1983) and Berger (1985). In these loss functions we can explicitly penalize highly dimensional parameter spaces to reflect a positive evaluation of parsimony.

Another way of penalizing large models is through the prior model probabilities. In particular, we can make $p(M_i)$ a decreasing function of l_i, the dimension of the model specific parameter vector η_i, e.g., we can assume $p(M_i) \propto 2^{-l_i}$. If, in addition, we can attach a particular status to one of the models, say M_i, we can follow a suggestion in Jeffreys (1961, p. 249, 253-254) to fix the prior probability of that model at a prespecified value, say $\frac{1}{2}$, irrespective of the number of models, m. While Jeffreys suggests to distribute the remaining prior probability evenly over the $m-1$ other models, we can also choose $p(M_i)$ for $i = 2, \dots, m$ depending on l_i, as explained above.

We wish to remind the reader that it is convenient to consider non-nested models, in order to avoid paradoxical situations where restrictions on the parameter space do not lead to a reduction in prior probability. We will come back to this in Section 5 and 7.

It is necessary to stress that posterior model probabilities are sensitive to prior specification. The role of prior model probabilities is evident from (3.3). For pairwise comparison, however, the influence of prior model probabilities can easily be isolated if we reason in terms of posterior odds which can be represented as

$$\frac{P(M_i|y)}{P(M_j|y)} = \frac{P(M_i)p_i(y)}{P(M_j)p_j(y)}, \tag{3.5}$$

i.e. the product of the prior odds, $P(M_i)/P(M_j)$, and the Bayes factor, $B_{ij} = p_i(y)/p_j(y)$. B_{ij} measures the relative within-sample predictive power of M_i and M_j and summarizes the data evidence in favour of M_i. By the very definition of the within-sample predictive density (3.4), its value for a given y is calculated through averaging the likelihood, $p_i(y|\omega_i)$, with the use of the prior density, $p_i(\omega_i)$, as the weighting function. Proper prior densities which are very flat relative to the likelihood will lead to much smaller values of $p_i(y)$ for a given data vector than prior densities which are similar to the likelihood in their location and spread. This is not very surprising, however, as it matters whether we give a lot of prior weight to areas with negligible likelihood values. In some cases, the posterior densities will also differ between these priors. But there are situation when the posterior density is unaffected, yet $p_i(y)$ varies enormously with the prior $p_i(\omega_i)$. Let us consider an example.

Assume that ω_i, is a real-valued parameter ($\Omega_i = \mathbb{R}$) and the likelihood is negligible outside some interval $[a, b]$, in the sense that

$$\int_{\Omega_i} p_i(y|\omega_i) d\omega_i \cong \int_a^b p_i(y|\omega_i) d\omega_i \equiv L_i. \tag{3.6}$$

For posterior inference on ω_i it is irrelevant whether we take $p_i(\omega_i) = \frac{1}{b-a} I(a \leq \omega_i \leq b)$, the uniform prior on $[a, b]$, or $p_i(\omega_i) = \frac{1}{2d+b-a} I(a - d \leq \omega_i \leq b + d)$, where $d > 0$. In both cases, the posterior density function will be just the normalized likelihood function, $p_i(\omega_i|y) \cong L_i^{-1} p_i(y|\omega_i)$, independent of d (with equality $d = 0$). The within-sample predictive value, however, will be a decreasing function of d,

$$p_i(y) = \int_{a-d}^{b+d} p_i(\omega_i) p_i(y|\omega_i) d\omega_i \cong \frac{1}{2d + b - a} L_i. \tag{3.7}$$

For any fixed $p_j(\omega_j)$, we can make the Bayes factor B_{ij} arbitrarily small just by broadening $[a - d, b + d]$. Thus, Bayes factors and posterior model probabilities strongly depend on our prior assumptions even if we restrict ourselves to proper prior density functions.

Moreover, improper prior structures cannot be used for model choice problems as widely and automatically as they are used for pure parameter estimation. For the latter, it is enough to specify an improper prior density kernel $g_i(\omega_i)$ such that

$$K_i = \int_{\Omega_i} p_i(y|\omega_i) g_i(\omega_i) d\omega_i < +\infty \tag{3.8}$$

as then the posterior density function

$$p_i(\omega_i|y) = K_i^{-1} g_i(\omega_i) p_i(y|\omega_i) \tag{3.9}$$

is well defined. Note that the same arbitrary constant, $c_i > 0$, appears in both $p_i(\omega_i) = c_i g_i(\omega_i)$ and $p_i(y) = c_i K_i$, and thus cancels in (3.9). If we calculate Bayes factor, however, the dependence of $p_i(y)$ on c_i implies that B_{ij} is defined only up to an indeterminate constant and, therefore, posterior model probabilities are undefined.

Several attempts have been made to overcome this problem, and the main motivation behind them is the desire to use noninformative (usually improper if Ω_i is not compact) priors in Bayesian analysis. Recently, Consonni and Veronese (1992) proposed a new method of determining Bayes factors under improper priors. Their method rests upon the idea that an improper prior may be regarded as a limit of proper priors defined on a sequence of (appropriately chosen) compact subsets converging to the whole parameter space. They propose a method to indirectly assign such a sequence, based on the imaginary training sample device as formalized by Spiegelhalter and Smith (1982). The idea of an imaginary data set was also used by Pettit (1992) and basically boils down to fixing the ratio of undefined constants c_i/c_j such that B_{ij} calculated on the basis of the smallest possible experiment we can use to distinguish two nested models is approximately 1.

We should mention the reference prior method of Bernardo (1979) and Berger and Bernardo (1992), which was applied to posterior odds testing and model choice by Bernardo (1980) and Pericchi (1984). The reference prior approach aims at deriving standard prior structures, justified by information theoretic arguments. Pericchi (1984) extends the use of a measure of expected information gain to the assessment of the prior model weights $p(M_i)$.

Another way of solving the problem of indeterminacy of Bayes factors under improper prior is to use prior densities of the form

$$p_i(\omega_i) = p_i(\theta, \eta_i) = p(\theta)p_i(\eta_i|\theta), \tag{3.10}$$

where $p(\theta) = cg(\theta)$ is the common improper prior on θ and $p_i(\eta_i|\theta)$ are proper $(i\ = 1, \dots, m)$. Note that now all $p_i(y)$ depend on the same arbitrary constant $c > 0$ which cancels in (3.3) and (3.5), leading to uniquely defined Bayes factors and posterior model probabilities. If, however, the prior structure in (3.10) does not correspond to strongly held prior opinions, checking the sensitivity of Bayes factors to the form of $p_i(\eta_i|\theta), i = 1, \dots, m$, is very important.

Although the posterior odds method is fundamental, there are other Bayesian approaches to model comparison. The so-called Lindley type or highest posterior density (HPD) testing can be used for pairwise comparison with the common special case (the “smallest” model) or with some general model, nesting the rest. In order to explain this approach, we assume that there are two models, M_1, with no model specific parameters, and M_2, with $\eta \in H_2$ as its model specific parameter. Moreover, we assume that M_1 is nested in M_2, which reduces to M_1 iff $\eta = \eta^*$. Note that our previous assumption of non-nested models would mean that $\eta^* \notin H_2$, while now that is no longer necessary since we do not attach probabilities to models (i.e. we do not put any prior mass at $\eta = \eta^*$). Within the Lindley type approach, model comparison is based on the marginal posterior density of $\eta, p_2(\eta|y)$, and amounts to checking whether η^* lies in some HPD region with a preassigned posterior probability content. If not, M_1 is rejected. For $0 < \alpha < 1$, the $1 - \alpha$ HPD region for η is the subset of H_2 given by

$$C = \{\eta \in H_2 : \ p_2(\eta|y) \geq k(\alpha)\} \tag{3.11}$$

where $k(\alpha)$ is the largest constant that satisfies

$$\int_C p_2(\eta|y)d\eta \geq 1 - \alpha.$$

This procedure minimizes the volume of C for a given probability content and could lead to disjoint intervals, e.g. indicating that prior and sample information are not in accordance with each other.

This testing procedure is asymmetric in that M_1 has to correspond to a restricted version of M_2. Under noninformative prior, this approach often provides us with a direct Bayesian interpretation of sampling theory testing procedures, as will be discussed in Section 5.

Yet another approach to model comparison, which has been developed very recently, is a Bayesian version of model encompassing. The basic idea there is that a desirable characteristic of a model lies in the ability to "explain" the inference obtained from other models. For a thorough exposition, see Florens and Mouchart (1994) and Florens et al. (1991). In order to verify whether M_1 can encompass M_2, we need to reinterpret the parameters of M_2 within the context of M_1. This is conceptually easy within a Bayesian framework by specifying a conditional transition density $p_1(\omega_2|\omega_1)$, also called a Bayesian pseudo-true value. The latter can be thought of as a Bayesian counterpart to the classical notion of pseudo-true value where an expectation or probability limit (under M_1 of a statistic which "naturally" appears in M_2) is used to express ω_2 as a deterministic function of ω_1.

This fact allows us to derive a posterior density for ω_2, the parameters in M_2, on the basis of M_1:

$$p_1(\omega_2|y) = \int_{\Omega_1} p_1(\omega_1|y)p_1(\omega_2|\omega_1)d\omega_1. \tag{3.12}$$

Comparison between this derived posterior for ω_2 using M_1 and its actual posterior under $M_2, p_2(\omega_2|y)$, will now tell us whether M_1 can account for the results obtained with M_2, i.e. whether M_1 encompasses M_2. Typically, we will focus on a parameter of interest $\varphi = f(\omega_2)$ and examine the discrepancy between the induced posterior densities $p_1(\varphi|y)$ and $p_2(\varphi|y)$ or between characteristics (e.g. moments) of these densities. The main problem in this testing procedure lies in the specification of the transition probability characterised by $p_1(\omega_2|\omega_1)$. Florens et al. (1991) and Florens and Mouchart (1994) suggest some possible strategies for choosing $p_1(\omega_2|\omega_1)$, but an "automatic" procedure does not seen to exist for the construction of transition probabilities. Clearly, the conclusions of this testing procedure will depend on the particular $p_1(\omega_2|\omega_1)$ chosen.

Originally motivated by the development of parsimonious modelling strategies, the concept of encompassing is inherently asymmetric. The parameter of interest is defined in terms of the parameters of M_2 and the question in often whether M_1, a simpler model or a model proposed by another researcher, can explain the results obtained with "your" model M_2. If not, this casts doubt upon M_1 but it does not necessarily lend credibility to M_2.

It should be clear to the reader that all these methods of testing explicitly refer to an alternative hypothesis. Pure significance testing, where an alternative is not specified, can not be justified from a formal Bayesian perspective, and is generally not accepted by Bayesians. See Poirier (1992) for a discussion and Hodges (1990) for a different point of view.

Finally, the HPD approach relies upon a nested structure of the contending models. Both posterior odds and encompassing procedures can easily deal with nonnested models.

4 Pooling inferences from individual models

If we are interested in testing competing theories or rival models, then model choice procedures are naturally unavoidable. Formally speaking, we aim at estimating the model label, which can be treated as a discrete parameter, and then we conduct inference conditionally on its estimate. This two-step strategy is usually called pretesting. However, when our research goals are different from testing theories, we can take into account all our specification uncertainty and average inferences over competing models.

If our interest is only in estimating the parameters common to all models, or in predicting observables, we can avoid selecting a particular model and use a Bayesian pooling approach. It amounts to computing the weighted average of posterior or out-of-sample predictive densities with posterior model probabilities as weights. Therefore, the fully marginal posterior and predictive density functions for the common parameters and observables are, respectively,

$$p(\theta|y) = \sum_{i=1}^{m} p(M_i|y) p_i(\theta|y) \tag{4.1}$$

and

$$p(y_f|y) = \sum_{i=1}^{m} p(M_i|y) p_i(y_f|y), \tag{4.2}$$

where $p(M_i|y)$ are the posterior model probabilities given by (3.3), and $p_i(\theta|y)$ and $p_i(y_f|y)$ are the model specific posterior and predictive density functions. Note that now in (4.2) both parameter uncertainty (as captured by the posterior distribution of the ω_i's) and specification uncertainty (as captured by (3.3) within the class of models considered) are entirely taken into account.

Other formulations of this Bayesian pooling strategy, in terms of model mixtures, are given in Osiewalski and Steel (1993a,b). A recent microeconomic application is provided by van den Broeck et al. (1994) who consider stochastic cost frontier models which differ in the distribution of the inefficiency term.

Remark that both the model choice approach and the pooling strategy can be given a decision theoretic motivation. If the assumed loss function only involves the model label, $\lambda \in \Lambda = \{1, \dots, m\}$, and its estimate $\hat{\lambda} \in \Lambda$, model choice is based on the posterior model probabilities in (3.3). Formally, if $L(\lambda, \hat{\lambda})$ denotes such a loss function, where $\hat{\lambda}$ is the decision, posterior expected loss is

$$\begin{aligned} E\{L(\lambda, \hat{\lambda})|y\} &= \sum_{i=1}^{m} L(i, \hat{\lambda}) p(\lambda = i|y) \\ &= \sum_{i=1}^{m} L(i, \hat{\lambda}) p(M_i|y), \end{aligned} \tag{4.3}$$

and one chooses that $\hat{\lambda} = j$ which minimizes (4.3).

On the other hand, if the loss structure only depends on observables or parameters common to all models, then calculating posterior expected loss automatically entails mixing over models. For example, if we wish to estimate θ by $\hat{\theta}$, then posterior expected loss is

$$E\{L(\theta,\hat{\theta})|y\} = \int_{\Theta} L(\theta,\hat{\theta})p(\theta|y)d\theta, \tag{4.4}$$

where $p(\theta|y)$ is the mixture of individual posterior densities in (4.1). To date, we have not been able to find a loss function involving both λ and θ that would lead to conducting inference on θ solely on the basis of one model. To us this seems to indicate that advocates of pretesting are not necessarily deriving their motivation from formal decision theory.

Alternatively, forecasting observables can be of interest, as in Min and Zellner (1993), Palm and Zellner (1992), and Zellner, Hong and Gulati (1990). In the particular case of predictive squared error loss, Min and Zellner (1993) confirm the general result that mixing always leads to optimal forecasts, provided the set of models we consider is exhaustive. The latter assumption is implicitly maintained throughout the present paper. Min and Zellner (1993) stress that if this assumption does not hold, mixing need not be the preferred strategy.

5 Discriminating between the means in linear regression models

In this section we shall focus on a case which has attracted a lot of attention in the literature. This, so-called, choice of regressors problem is usually cast in a nested framework, in which case HPD testing can be conducted. Of course, posterior odds procedures can deal with both nested and non-nested environments. We shall explain matters here in the context of nested models for HPD testing and non-nested models for the posterior odds and encompassing procedures.

Assume that $m = 2$,

$$M_1 : y = X\beta + \varepsilon \tag{5.1}$$

and

$$M_2 : y = X\beta + Z\eta + \varepsilon, \tag{5.2}$$

where ε has an n-variate Normal distribution with mean 0 and covariance matrix $\sigma^2 I_n$; $\sigma \in \mathbb{R}_+$, $\beta \in \mathbb{R}^k$ and $\eta \in \mathbb{R}^l$ are unknown parameters; X and Z are matrices of (weakly) exogenous explanatory variables, and $W = [X\ Z]$ is of full column rank, $k + l$.

This is a special case of our general framework, where $\theta = (\beta'\sigma)'$ groups the common parameters, $\omega_1 = \theta$ and thus $H_1 = \emptyset$, $\omega_2 = (\theta'\eta')'$ and $H_2 = \mathbb{R}^l$ and M_1 can be obtained from M_2 by imposing the restriction $\eta = 0$. This means that M_1 is nested in M_2 unless that particular point, i.e. $\eta^* = 0$, is excluded from H_2.

First, we focus on a Lindley type test for the restriction $\eta = 0$. Assume an improper uniform prior of $(\beta, \eta, \ln(\sigma)) \in \mathbb{R}^{k+l+1}$, which corresponds to

$$p_2(\beta, \eta, \sigma) \propto \sigma^{-1}. \tag{5.3}$$

As is well known, the marginal posterior distribution of $(\beta'\eta')'$ is the $(k + l)$-variate Student t distribution with $\nu = n - (k + l)$ degrees of freedom, location vector

$$\begin{bmatrix} \hat{\beta} \\ \hat{\eta} \end{bmatrix} = (W'W)^{-1}W'y \tag{5.4}$$

and precision matrix $s^{-1}W'W$, where

$$s^2 = \frac{1}{\nu}\left(y - X\hat{\beta} - Z\hat{\eta}\right)'\left(y - X\hat{\beta} - Z\hat{\eta}\right). \tag{5.5}$$

The marginal posterior distribution of η alone is the l-variate Student t distribution with ν degrees of freedom, location vector $\hat{\eta}$ and precision matrix $A = s^{-2}Z'M_XZ$, where $M_X = I_n - X(X'X)^{-1}X'$. Since this is a unimodal ellipsoidal distribution, HPD regions for η are the convex sets bounded by isodensity ellipsoids, i.e. the sets $R(u_0) = \{\eta^* \in \mathbb{R}^l : u(\eta^*) \leq u_0\}$, where $u(\eta^*) = (\eta^* - \hat{\eta})'A(\eta^* - \hat{\eta})$ and $u_0 \in \mathbb{R}_+$ corresponds to the largest ellipsoid (i.e. to the minimal density level) within the HPD region. The posterior probability that η lies within $R(u_0)$ is

$$\begin{aligned} P\{\eta \in R(u_0)|\text{data}\} &= P\{u(\eta) \leq u_0|\text{data}\} \\ &= P\{F_{l,\nu} \leq \frac{1}{l}u_0\}, \end{aligned} \tag{5.6}$$

because the posterior distribution of the quadratic form $\frac{1}{l}u(\eta)$ is the F distribution with l and ν degrees of freedom, see e.g. Zellner (1971, p. 385). If we define the critical value F_α by the equation $P\{F_{l,\nu} > F_\alpha\} = \alpha$, then $\eta^* = 0$ lies on the boundary of the HPD region of posterior probability $1-\alpha$ iff $\frac{1}{l}u(0) = F_\alpha$, lies within region iff $\frac{1}{l}u(0) < F_\alpha$, and is outside this region iff $\frac{1}{l}u(0) > F_\alpha$. Of course, the value chosen for α is essentially arbitrary.

Note that $\frac{1}{l}u(0) = \frac{1}{l}\hat{\eta}'A\hat{\eta}$ is the sampling-theory F statistic for testing $H_0 : \eta = 0$ versus $H_1 : \eta \neq 0$. Thus, under the improper prior structure (5.3), the approach described above gives us a direct Bayesian interpretation of the classical F test in terms of HPD regions for η.

For a posterior odds approach we assume that the prior is proper on η, as we discussed before. If we choose a prior of the natural-conjugate type, we obtain for M_2 which now excludes $\eta^* = 0$:

$$p_2\left(\begin{matrix}\beta\\ \eta\end{matrix}\middle|\sigma^2\right) = f_N^{k+l}\left(\begin{matrix}\beta\\ \eta\end{matrix}\middle|\gamma_0, \sigma^2 N_0^{-1}\right) \tag{5.7}$$

a $(k+l)$-variate Normal density function with mean γ_0 and covariance matrix $\sigma^2 N_0^{-1}$, and

$$p_2(\sigma^2) = f_{i\gamma}(\sigma^2|s_0, \nu_0), \tag{5.8}$$

an inverted gamma density with ν_0 degrees of freedom and mean $s_0/(\nu_0 - 2)$ if $\nu_0 > 2$, corresponding to a gamma density on σ^{-2} with hyperparameters $\nu_0/2$ and $s_0/2$ and mean ν_0/s_0.

All calculations can now be done analytically, and we can show [see e.g. Kiefer and Richard (1987)] that if we take[2] $p_1(\beta, \sigma^2) = p_2(\beta, \sigma^2|\eta = 0)$ we obtain

[2] This reflects that the interpretation of $\theta = (\beta, \sigma^2)$ is the same under both models

$$B_{12} = \frac{\Gamma(\frac{\nu_0 + l + n}{2})\Gamma(\frac{\nu_0}{2})s_0^{\frac{1}{2}l}|I_n - X\,N_{11}^{-1}X'|^{\frac{1}{2}}}{\Gamma(\frac{\nu_0 + n}{2})\Gamma(\frac{\nu_0 + l}{2})|I_n - W\,N^{-1}W'|^{\frac{1}{2}}}$$

$$\times \frac{\{s_0 + (y - X\beta_0)'(I_n - X\,N_{11}^{-1}X')(y - X\beta_0)\}^{-\frac{\nu_0+l+n}{2}}}{\{s_0 + (y - X\beta_0)'(I_n - W\,N^{-1}W')(y - X\beta_0)\}^{-\frac{\nu_0+n}{2}}}, \tag{5.9}$$

Where $N = N_0 + W'W$ is assumed nonsingular with N_{11} its $k \times k$ upper left block and we have chosen $\gamma_0 = (\beta_0'0')'$ corresponding to the partitioning of W.

If s_0 and N_0 are relatively small and $\beta_0 = 0$, then the quantities in curly brackets in (5/9) are approximately the residual sums of squares after regressing y on X in M_1 and on W in M_2. This provides a link with classical F tests, albeit with a rather different interpretation.

If we are not willing to specify a prior as in (5.7) – (5.8), a simple strategy to calculate the posterior odds for M_1 and M_2 is to specify a uniform improper prior for $(\beta, \ln(\sigma)) \in \mathbb{R}^{k+1}$, but a proper (conditional) prior density $p_2(\eta|\theta)$ over $H_2 = \mathbb{R}^l\{0\}$. Furthermore we continue to exclude $\eta^* = 0$ from H_2 and attach the prior probability mass $P(M_1)$ to this particular point. Thus, the prior density adopted under this strategy is of the form (3.10). The data evidence in favour of M_1, summarized by the Bayes factor B_{12}, will be strongly influenced by $p_2(\eta|\theta)$, and, therefore, a careful choice of this prior density is required.

Building on the earlier work of Jeffreys (1961, Ch. V), Zellner and Siow (1980, p. 593-594) suggested for $p_2(\eta|\theta)$ the l-variate Cauchy density with zero location vector and precision matrix $\frac{1}{n}\sigma^{-2}Z'M_XZ$, a matrix independent of β and suggested by the form of the information matrix. Under this Cauchy prior of η given θ and the common improper prior of $\theta = (\beta'\sigma)'$,

$$p(\beta, \sigma) \propto \sigma^{-1}, \tag{5.10}$$

Zellner and Siow (1980) derive the following approximate expression for the Bayes factor

$$B_{12} \cong \frac{\sqrt{\pi}}{\Gamma\left(\frac{l+1}{2}\right)}\left(\frac{\nu}{2}\right)^{\frac{l}{2}}(1 + \frac{1}{\nu}\hat{\eta}'A\hat{\eta})^{-(\nu-1)/2} \tag{5.11}$$

If ν is large, the last factor in (5.11) can be further approximated by its limit, $\exp\left(-\frac{1}{2}\hat{\eta}'A\hat{\eta}\right)$. As $\frac{1}{l}\hat{\eta}'A\hat{\eta}$ is the usual sampling-theory F statistic, the latter approximation shows that any "critical value" for rejecting M_1 should be a function of $\nu = n - (k + l)$, increasing with $\ln(\nu)$.

In order to explain the encompassing testing procedure in simple terms, we shall assume that σ^2 is known. To avoid notational problems, let us rewrite M_1 as

$$M_1 : y = X\delta + \varepsilon \tag{5.12}$$

to stress that β in M_2 and δ in M_1 are not necessarily the same. Prior densities will now be chosen compatible with (5.7), namely

$$p_1(\delta) = f_N^k(\delta|\delta_0, \sigma^2 M_0^{-1}) \tag{5.13}$$

$$p_1(\gamma) = f_N^{k+l}(\gamma|\gamma_0, \sigma^2 N_0^{-1}) \tag{5.14}$$

where $\gamma = (\beta'\eta')'$. This leads to the following posterior densities:

$$p_1(\delta|y, X) = f_N^k(\delta|\delta_*, \sigma^2 M^{-1}) \tag{5.15}$$

$$p_2(\gamma|y, W) = f_N^{k+l}(\gamma|\gamma_*, \sigma^2 N^{-1}) \tag{5.16}$$

with $M = M_0 + X'X$, N as defined after (5.9), and

$$\begin{aligned} \gamma_* = (\beta_*'\eta_*')' &= N^{-1}(N_0\gamma_0 + W'y) \\ \delta_* \qquad\quad &= M^{-1}(M_0\delta_0 + X'y). \end{aligned} \tag{5.17}$$

Consider now only the class of Normal transition densities:

$$p_1(\gamma|\delta) = f_N^{k+l}(\gamma|r + R\delta, \sigma^2 V), \tag{5.18}$$

which implies that the derived posterior density of γ under M_1 now becomes:

$$p_1(\gamma|y,X) = f_N^{k+l}(\gamma|r + R\delta_*, \sigma^2(V + RM^{-1}R')). \tag{5.19}$$

For a particular parameter of interest $\varphi = f(\gamma) \in \Phi$, the discrepancy between $p_2(\varphi|y,W)$ derived from (5.16) and $p_1(\delta|y,X)$ from (5.19) can now be measured by e.g. the negative entropy (or Kullback-Leibler divergence):

$$D_\varphi(y,W) = \int_\Phi ln \frac{p_1(\varphi|y,X)}{p_2(\varphi|y,W)} p_1(\varphi|y,X) d\varphi \tag{5.20}$$

In the case that $\varphi = \beta$, this measure would become:

$$\begin{aligned} D_\beta(y,W) &= \frac{1}{2}\{-ln|V_{11} + R_1M^{-1}R_1'| - ln|N_{11.2}| + trN_{11.2} \\ &\times (V_{11} + R_1M^{-1}R_1') + \sigma^{-2}(r_1 + R_1\delta_* - \beta_*)'N_{11.2}(r_1 + R_1\delta_* - \beta_*) - k\}, \end{aligned} \tag{5.21}$$

where r_1 and R_1 are the first k elements and rows of r and R, V_{11} is the $k \times k$ upper left block of V and $N_{11.2} = N_{11} - N_{12}N_{22}^{-1}N_{21}$ if we partition $N = \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix}$ conformably.

A possible choice for the transition probability in (5.18) would be a degenerate distribution at $\gamma = (\delta'0')'$, indicating that one interprets β and δ as the same parameter. In that case, $r = 0$, $R = (I_k\ 0)'$ and $V = 0$, such that we obtain

$$\begin{aligned} D_\beta(y,W) &= \frac{1}{2}\{\ln|M| - \ln|N_{11.2}| + trN_{11.2}M^{-1} + \sigma^{-2}(\delta_* - \beta_*)'N_{11.2} \\ &(\delta_* - \beta_*) - k\}. \end{aligned} \tag{5.22}$$

If we identify δ with β, a coherent prior specification implies that $p_1(\delta) = p_2(\beta|\eta = 0)$ (see also footnote 2), and (5.22) becomes a Bayesian version of the Hausman (1978) test, with e.g.

$$\gamma_0 = \binom{\delta_0}{0}$$

$$M_0 = (N_0)_{11}$$

Where $(N_0)_{11}$ is defined analogously to N_{11}. See Florens and Mouchart (1994). In the latter case

$$\delta_* - \beta_* = N_{11}^{-1}N_{12}\eta_*$$

and

$$p_1(\beta|y,X) = f_N^k(\beta|\delta_*,\sigma^2 M^{-1}) = p_2(\beta|y,W,\eta = 0);$$

i.e., due to the coherent prior structure, the derived posterior of β under M_1 coincides with the conditional posterior of β given $\eta = 0$ under M_2. In case β and η are also uncorrelated (and thus independent in this Normal framework) in the posterior under M_2 (5.16), we have $N_{12} = 0$ and thus $N_{11.2} = M$, as well as $\delta_* = \beta_*$, so that $D_\beta(y,W) = 0$. Clearly, in that case M_1 leads to exactly identical inference on β as M_2 and thus encompasses M_2 for β with the degenerate transition probability and a coherent prior in (5.13)-(5.14).

Florens et al. (1991) suggest finding an "optimal" transition probability by minimizing the expectation of $D_\varphi(y,W)$ with respect to the within-sample predictive density under M_1. If we now choose $\varphi = \gamma$, such optimality will be achieved for

$$r = N^{-1}(N_0\gamma_0 - W'X(X'X)^{-1}M_0\delta_0) \tag{5.23}$$

$$R = N^{-1}W'X(X'X)^{-1}M \tag{5.24}$$

$$V = N^{-1} - RM^{-1}R', \tag{5.25}$$

provided the latter is semipositive definite, leading to

$$D_\gamma(y,W) = \frac{1}{2\sigma^2} y'M_X Z N_{22.1}^{-1} Z'M_X y, \tag{5.26}$$

where $M_X = I_n - X(X'X)^{-1}X'$ and $N_{22.1} = N_{22} - N_{21}N_{11}^{-1}N_{12}$.
Florens amd Mouchart (1994) note the asymptotic similarity of (5.26) to a Wald test statistic. Indeed, for large n the influence of the prior can be neglected, so that $N_{22.1}$ will be approximated by $Z'M_X Z$, and $D_\gamma(y,W)$ will be close to[3] $\frac{1}{2}\hat{\eta}'A\hat{\eta} = \frac{1}{2}u(0)$ used

[3] In A the quantity s^2 defined in (5.5) constitutes a consistent estimator of σ^2, which was assumed known for explaining the encompassing approach.

in HPD testing in (5.6). This is the sampling-theory F statistic times a factor $l/2$. Operational calibration of $D_\gamma(y, W)$ still seems an open question. The within-sample predictive density under M_1, given by

$$p_1(y) = f_N^n(y|X\delta_0, \sigma^2(I + XM_0^{-1}X'))$$

could be a useful guide in practice.

Using the same optimal values for r, R and V as in (5.23)-(5.25), we obtain

$$\begin{aligned} D_\beta(y, W) &= \frac{1}{2\sigma^2} y' M_X Z N_{22}^{-1} N_{21} N_{11.2}^{-1} N_{12} N_{22}^{-1} Z' M_X y \\ &= D_\gamma(y, W) = \frac{1}{2\sigma^2} y' M_X Z N_{22}^{-1} Z' M_X y. \end{aligned} \tag{5.27}$$

Regardless of the choice of the prior hyperparameters δ_0, γ_0, M_0 and N_0, the discrepancy $D_\beta(y, W) = 0$ whenever β and η are posterior independent under this “optimal” Normal transition density.

An alternative approach to variable selection is developed in George and McCulloch (1993). They specify the prior on the elements of η as consisting of a mixture of two Normal distributions, one with a large variance (corresponding to retaining that variable) and one very spiked around zero (corresponding to deleting variable).[4] Latent variables identify the choice between both types of distributions. Inference on these latent variables thus relates directly to variable selection, and the computational feasibility of this approach derives from the use of Gibbs sampling, a Markov chain simulation method described in Gelfand and Smith (1990) and in an introductory way in Casella and George (1992).

6 Dynamic regression models under competing correlation structures

In this section we compare m dynamic regression models (each with q lagged dependent variables) that differ only in their correlation structure. In addition, we allow for general elliptical distributions of the error vector. Bayesian posterior analysis is not fundamentally affected by the dynamic character of the model, and posterior odds are obtained in

[4] However, they avoid putting prior mass on $\eta = 0$, unlike the “spike and slab” mixtures of Mitchell and Beauchamp (1988).

the same fashion as in Osiewalski and Steel (1993a) who treat the static case. Indeed, posterior results are based on the likelihood function, the functional form of which is not changed by introducing dynamics. Within a Bayesian framework, the latter will only complicate prediction [see Chow (1973) and Koop, Osiewalski and Steel (1994)].

In this case, the structure of the problem often imposes nonnested testing, which rules out HPD procedures and greatly complicates classical approaches.[5]

We consider m dynamic linear regression models $(i = 1, \dots, m)$

$$M_i : y = Y_{-1}\alpha + X\beta + \varepsilon_i \tag{6.1}$$

where Y_{-1} is an $n \times q$ matrix containing lagged values of n dimensional vector y as well as the necessary initial values y_0, and X groups k other weakly exogenous variables. The error vector ε_i is assumed to have an n-variate elliptical distribution with location vector 0 and dispersion matrix $\sigma^2 V_i$, with σ^2 a common scale factor, and $V_i = V_i(\eta_i)$ a model specific PDS matrix function of the l_i dimensional η_i. The m models thus only differ in the structure of V_i.

For national convenience, let $Z = (Y_{-1}\ X)$ and $\gamma' = (\alpha'\ \beta')$. As a result of the unitary Jacobian of the transformation from ε to y, the data density corresponding to M_i is:

$$p_i(y|y_0, X, \gamma, \sigma^2, \eta_i) = (\sigma^2)^{-\frac{n}{2}} |V_i|^{-\frac{1}{2}} g_i[(y - Z\gamma)'\sigma^{-2}V_i^{-1}(y - Z\gamma)]. \tag{6.2}$$

In (6.2) the nonnegative function $g_i[\cdot]$ is such that $u^{\frac{n}{2}-1}g_i(u)$ is integrable in $\mathbb{R}_+$, $i = 1, \dots, m$; see Dickey and Chen (1985). This general class covers many specific multivariate densities, like Normal, Student t or Pearson II. Due to the linearity of the transformation from ε to y, the data density still belongs to the elliptical class. Finally, the entire analysis will be conducted conditionally upon y_0. For alternative treatments of initial values see e.g. Zellner (1971) and Richard (1979).

In order to complete the Bayesian model, we specify a prior density as in (3.10) on the parameters of M_i:

[5] Sampling theory procedures boil down to point optimal tests, as described e.g. in King (1983, 1987-1988) and Inder (1990). All models are restated in terms of simple hypotheses by conditioning on particular values of α and η_i in (6.1).

$$p_i(\gamma, \sigma^2, \eta_i) = c_1 \sigma^{-2} p(\gamma) p_i(\eta_i), \tag{6.3}$$

a product of the usual improper prior on σ^2, a prior on the common coefficients γ, and a proper prior on η_i, with $c_1 > 0$. The Jeffreys' type prior on σ^2 can be shown, as in Osiewalski and Steel (1993b), to lead to exactly the same joint density of (y, γ, η_i) as under Normality of the disturbances in (6.1), namely

$$p_i(y, \gamma, \eta_i | y_0, X) = c_1 \Gamma\left(\frac{n-k-q}{2}\right) \pi^{-\frac{n-k-q}{2}} p(\gamma) p_i(\eta_i)$$
$$h_i(\eta_i) f_S^{k+q}\left(\gamma \middle| n-k-q, \hat{\gamma}_i, \frac{n-k-q}{SSE_i} Z' V_i^{-1} Z\right), \tag{6.4}$$

with

$$h_i(\eta_i) = |V_i|^{-\frac{1}{2}} |Z' V_i^{-1} Z|^{-\frac{1}{2}} (SSE_i)^{-\frac{n-k-q}{2}}, \tag{6.5}$$

and the $(k+q)$-variate Student t density appearing in (6.4) has $n-k-q$ degrees of freedom, location vector $\hat{\gamma}_i = (Z' V_i^{-1} Z)^{-1} Z' V_i^{-1} y$ and the precision matrix involves $SSE_i = (y - Z\hat{\gamma}_i)' V_i^{-1} (y - Z\hat{\gamma}_i)$; finally, we implicitly assume Z to be of full column rank.

Clearly, γ can be integrated out analytically from (6.4) if we assume an improper uniform prior in (6.3)

$$p(\gamma) = c_2. \tag{6.6}$$

This convenient case will be treated here in some detail, whereas for independent Student t priors on γ the results in Osiewalski and Steel (1993a) can easily be adapted. Remark that in the context of dynamic models the choice of (6.6) does not exclude nonstationarity of the process for y. Imposing stationarity requires restricting the parameter space of α, which would add q dimensions to the numerical integration in the sequel.

Under M_i, the use of (6.3) and (6.6) leads to the Student t conditional posterior of γ, given η_i, implicit in (6.4), and the following marginal posterior of η_i:

$$p_i(\eta_i|y, y_0, X) = K_i^{-1} h_i(\eta_i) p_i(\eta_i), \tag{6.7}$$

Where we assume $K_i = \int h_i(\eta_i) p_i(\eta_i) d\eta_i$ to be finite, $i = 1, \ldots, m$. Evaluating K_i only requires l_i dimensional numerical integration. Assigning prior probability $p(M_i)$ to the i-th model, the posterior probability is now given by

$$p(M_i|y, y_0, X) = \frac{p(M_i)K_i}{\sum_{j=1}^{m} p(M_j)K_j}, \tag{6.8}$$

since the (improper) predictive densities are $p_j(y|y_0, X) = cK_j$ with the same constant c for all $j = 1, \ldots, m$. The Bayes factor B_{rs} for comparing M_r and M_s is equal to K_r/K_s, leading to the posterior odds $[p(M_r)/p(M_s)] \times B_{rs}$. Note that B_{rs} could take any value if we would allow $p_i(\eta_i)$ in (6.7) to be improper.

If the loss structure penalizes all incorrect decisions equally heavily, the Bayesian pretest procedure amounts to choosing the model with highest posterior model probability. In order to avoid pretesting, we can use mixtures of data densities, as explained in Section 4 [see Osiewalski and Steel (1993a)].

An approximate Lindley type procedure for testing autoregressive (AR) disturbances (M_2) against uncorrelated disturbances (M_1) is given in Bauwens and Rasquero (1993). They define the “Bayesian residuals” $\hat{\varepsilon}_1$ in (6.1) by replacing $\gamma' = (\alpha' \; \beta')$ by its posterior mean $\hat{\gamma}_1$ under $V_1 = I_n$ and the prior in (6.3) and (6.6). Replacing the unobservable error terms by these Bayesian residuals, they add p lagged values of $\hat{\varepsilon}_1$ in an effort to capture an AR(p) structure. An F test like in Section 5 is then conducted to examine whether $\eta_2 = 0$. Of course, the translation to the choice of regressors problem gives us a very simple test, but at the cost of a rather ad-hoc approximation. Section 7 will discuss this in the framework of an example.

The posterior odds procedure for testing M_i with $V_i \neq I_n$ against M_1 is approximated through the method of Laplace [see Tierney and Kadane (1986)] in Poirier (1988a), leading to a Bayesian analog of a score or Lagrange multiplier test.

Classical point-optimal testing procedures as in Inder (1990) can be interpreted as based on conditional Bayes factors, where we fix values for α and η_i rather than integrate them out with the prior.

7 An empirical example: The influence of advertising on sales

The theoretical setup in Section 6 is now applied to the study of the effect of advertising outlay A_t on a firm's dollar sales S_t. A popular specification in the literature is:

$$S_t = \alpha S_{t-1} + \beta_1 + \beta_2 A_t + \varepsilon_t \tag{7.1}$$

which directly fits into (6.1). The properties of the error term in (7.1) are more contentious. The "brand loyalty model" of Houston and Weiss (1975) assumes a first-order autoregressive [AR(1)] structure on ε_t. If we call this model M_1, then

$$V_1^{-1} = (1 - \eta_1)^2 I_n + \eta_1 A - \eta_1^2 B \tag{7.2}$$

where $\eta_1 \in (-1, 1), B = \text{Diag}(1,0 \ldots 0,1)$ and A is a tridiagonal matrix with 2 on the main diagonal and -1 on the other diagonals.

An alternative model, say M_2, is characterized by first-order moving-average [MA(1)] behavior of ε_t. This implies for $\eta_2 \in (-1, 1)$

$$V_2 = (1 + \eta_2)^2 I_n - \eta_2 A. \tag{7.3}$$

Berndt (1991, p. 386) calls M_2 with the added restriction that $\alpha + \eta_2 = 0$ the "Koyck lingering effects" model.

As a third possibility, we take M_3 to be the simple model corresponding to $V_3 = I_n$.

The analysis in conducted on monthly data for the Lydia E. Pinkham Medicine Company, covering the period September 1954 until June 1960 ($n = 70$ observations).
The data were compiled by Palda (1964) and a very complete historical survey of research in this field is found in Berndt (1991, Ch. 8).

We assume elliptical data densities for all models, as in (6.2), and adopt the prior in (6.3). For the proper priors on η_i we shall choose Beta distributions over $(-1,1)$, i.e. for $i = 1,2$

$$p_i(\eta_i) = \frac{\Gamma(a+b)}{2^{a+b-1}\Gamma(a)\Gamma(b)}(1+\eta_i)^{a-1}(1-\eta_i)^{b-1}. \tag{7.4}$$

In order to assess the sensitivity of our results with respect to changes in the prior, we shall use $(0.5, 0.5)$, $(1,1)$ and $(2,2)$ for the hyperparameters (a,b). Choosing the same priors for η_1 and η_2[6] and using $P(M_i) = \frac{1}{3}, i = 1, \dots, 3$, we obtain the posterior model probabilities [see (6.8)] recorded in Table 1.

Table 1: Posterior model probabilities

	prior	posterior		
		$(a,b) = (0.5, 0.5)$	$(1,1)$	$(2,2)$
M_1	$\frac{1}{3}$	0.1736	0.2352	0.2291
M_2	$\frac{1}{3}$	0.5687	0.7053	0.5980
M_3	$\frac{1}{3}$	0.2577	0.0595	0.1729

As can be expected, the Bayes factors against M_3, the uncorrelated model, were most affected by changing (a,b) in the prior. Indeed, then the different integrating constants in (7.4) matter (remember that we have chosen $p(\eta_1) = p(\eta_2)$ in all cases here). The Bayes factor B_{12} of AR(1) versus MA(1) ranges only from 0.3053 to 0.3831 as (a,b) changes, whereas B_{13} goes from 0.6738 to 3.956. However, the posterior densities of η_1 or η_2 are very close in all three cases. In fact, they would be virtually indistinguishable in Figure 1, which shows the posterior densities of η_1 and η_2 for the uniform prior $[(a,b) = (1,1)]$. This is basically the same phenomenon as that caused by different supports for uniform priors described in Section 3. For given values of $\eta_i = \eta_i^*$, and $\eta_j = \eta_j^*$ we can calculate conditional Bayes factors $B_{ij}^* = h_i(\eta_i^*)/h_j(\eta_j^*)$, which are plotted in Figure 2 for different values of $\eta_1^* = \eta_2^*$ (the latter assumption is again made to facilitate the presentation).

Since we use the prior on η_i to average out $h_i(\eta_i)$ in computing K_i and thus Bayes factors through (6.8), the influence of the integrating constant in (7.4) becomes clear, unless when it happens to cancel [since $p_i(\eta_i) = p_j(\eta_j)$].

[6] this is by no means necessary but reduces the number of cases to be presented here

For all three priors, most posterior probability is attributed to M_2, the MA(1) model. However, all models are allocated a nonnegligible probability, so that it makes sense to mix over models for posterior inference on γ or for prediction (as discussed in Section 4). Table 2 presents information on the posterior moments of γ under mixing.

If we would pretest and take the MA(1) model in all cases, our inference, especially on α and β_1 would be rather different. Table 3 summarizes the relevant information. Note that taking the model uncertainty into account is reflected in a substantial increase in the standard deviations in Table 2, except for β_2, for which all models lead to virtually the same inference. As the prior on η_i is relatively flat in the areas of high likelihood, posterior moments are not much affected by changes in (a, b) over the range used here.

Table 2: Mixed Posterior Moments of γ

	$(a, b) = (0.5, 0.5)$		(1,1)		(2,2)	
	mean	s.dev.	mean	s.dev.	mean	s.dev.
α	0.6688	(0.1257)	0.6993	(0.1040)	0.6694	(0.1139)
β_1	0.2909	(0.1590)	0.2517	(0.1312)	0.2888	(0.1455)
β_2	0.2032	(0.0365)	0.2035	(0.0350)	0.2052	(0.0359)

Table 3: Posterior Moments of γ under Pretesting

	$(a, b) = (0.5, 0.5)$		(1,1)		(2,2)	
	mean	s.dev.	mean	s.dev.	mean	s.dev.
α	0.7363	(0.0899)	0.7288	(0.0896)	0.7172	(0.0901)
β_1	0.2076	(0.1114)	0.2162	(0.1119)	0.2298	(0.1135)
β_2	0.1988	(0.0336)	0.2001	(0.0336)	0.2020	(0.0337)

One could ask whether the high posterior probability of M_2 implies that the data support the “Koyck lingering effects” model. Here we should remember that the latter imposes the restriction $\alpha + \eta_2 = 0$. Posterior moments of $\alpha + \eta_2$ are presented in Table 4, along with moments of η_1 and η_2.

Table 4: Moments for η_i

	$(a,b) = (0.5, 0.5)$		$(1,1)$		$(2,2)$	
	mean	s.dev.	mean	s.dev.	mean	s.dev.
$p(\eta_1)$	0	(0.71)	0	(0.58)	0	(0.45)
$p_1(\eta_1 \vert y, y_0, X)$	-0.31	(0.14)	-0.31	(0.14)	-0.29	(0.13)
$p(\eta_2)$	0	(0.71)	0	(0.58)	0	(0.45)
$p_2(\eta_2 \vert y, y_0, X)$	-0.47	(0.71)	-0.45	(0.17)	-0.41	(0.16)
$p_2(\alpha + \eta_2 \vert y, y_0, X)$	0.27	(0.71)	0.28	(0.10)	0.30	(0.10)

Clearly, HPD testing of the restriction $\alpha + \eta_2 = 0$ on the basis of its first two moments and a Student t approximation will lead to rejection of the Koyck lingering effects model versus M_2 for all commonly used posterior probability levels. If we would be willing to specify a proper prior on α, a posterior odds test of this restriction could also be conducted by considering one more model, say, M_4, where $\alpha = -\eta_2$ in $(-1,1)$.

The simple "Bayesian residuals test" by Bauwens and Resquero (1993), described in Section 6, approximates the actual posterior density of η_1 under an improper uniform prior on η_1 by a Student t density with mean -0.50, standard deviation 0.18 and 66 degrees of freedom. The resulting F values for the restriction $\eta_1 = 0$ is 8.18 which provides substantial information in favour of an AR(1) model against its special case of uncorrelated errors. Since Figure 1 clearly shows there is no posterior mass close to the boundaries of $(-1,1)$ for η_1, the effect of the truncation will be negligible and we can compare the approximate posterior used here with the actual posterior described in Figure 1 and Table 4 $[(a,b) = (1,1)]$. It should be clear that the usual HPD problem of an arbitrary choice of a significance level is compounded here by an approximation error.

8 Concluding remarks

In this paper we have attempted to give an account of three Bayesian testing principles, and their implications for model selection. By no means does our account claim to constitute an exhaustive survey. However, we hope it adds to the understanding of Bayesian testing procedures, both standard and less standard ones. We stress the specific weaknesses and limitation of different procedures and give some indications as to their applicability.

Two leading examples were chosen to illustrate the principles presented, both in the context of linear regression models. The choice of regressors problem focusses on the means of the observables to be modeled. All three approaches to this problem are discussed at some length in Section 5. Asymptotically, the same quantities often appear in the relevant expressions, but calibration of the tests may be very different. In that respect, the posterior odds test seems to have a great advantage, in that it leads to directly interpretable results and blends in perfectly with a formal decision analysis. In cases where a posterior odds test can be applies, we would induce the applied researcher to choose this approach.

That posterior odds testing can deal with rather complicated situations without excessive analytical requirements is illustrated in Section 6, where the choice between correlation structures in dynamic linear models is examined. The empirical application in Section 7 shows that this problem poses no computational difficulties and easily leads to addressing questions that are of relevance to applied researchers.

Acknowledgements: we are grateful to Carmen Fernández for many useful comments. The second author was partially funded by the CentER for Economic Research, Tilburg University.

References

Bauwens, L. and A. Rasquero (1993), “Approximate HPD Regions for Testing Residual Autocorrelation Using Augmented Regressions”, in W. Härdle and L. Simar (eds.), *Computer Intensive Methods in Statistics*, Physica-Verlag, Heidelberg, 47-61.

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

Berger, J.O. and J.M. Bernardo (1992), “On the Development of the Reference Prior Method”, in J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 4*, Oxford University Press, Oxford, 35-60 (with discussion).

Bernardo, J.M. (1979), “Reference Posterior Distributions for Bayesian Inference”, *Journal of the Royal Statistic Society*, B 41, 113-147 (with discussion).

Bernardo, J.M. (1980), “A Bayesian Analysis of Classical Hypothesis Testing”, in J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (eds.), *Bayesian Statistics*, Valencia University Press, Valencia, 605-618.\

Berndt, E.R. (1991), *The Practice of Econometrics: Classic and Contemporary*, Addison-Wesley, Reading.

Casella, G. and E. George (1992), “Explaining the Gibbs Sampler”, *The American Statistician*, 46, 167-174.

Chow, G.C. (1973), “Multiperiod Predictions from Stochastic Difference Equations by Bayesian Methods”, *Econometrica*, 41, 109-118 and 796 (erratum).

Consonni, G. and P. Veronese (1992), “Bayes Factors for Linear Models and Improper Priors”, in J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 4*, Oxford University Press, Oxford, 587-594.

Dickey, J.M. and C.H. Chen (1985), “Direct Subjective Probability Modelling Using Ellipsoidal Distributions”, in J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M.

Smith (eds.), *Bayesian Statistics 2*, North-Holland, Amsterdam, 157-182 (with discussion).

Florens, J.P. D.F. Hendry and J.F. Richard (1991), “Encompassing and Specificity”, GREMAQ Discussion Paper 91c, Université de Toulouse, Toulouse.

Florens, J.P. and M. Mouchart (1994), “Bayesian Testing and Testing Bayesians”, in G.S. Maddala, C.R. Rai and H.D. Vinod (eds.), *Handbook of Statistics, Vol. 11: Econometrics*, North-Holland, Amsterdam.

Gaver, K.M. and M.S. Geisel (1974), “Discriminating Among Alternative Models: Bayesian and non-Bayesian Methods”, in P. Zarembka (ed.), *Frontiers of Econometrics*, Academic Press, New York, 49-77.

Gaver, K.M. and M.S. Geisel (1976), “Discriminating Among Linear Models with Interdependent Disturbances”, *Econometrica*, 44, 337-343.

Gelfand, A.E. and A.F.M. Smith (1990), “Sampling Based Approaches to Calculating Marginal Densities”, *Journal of the American Statistical Association*, 85, 398-409.

George, E.I. and R.E. McCulloch (1993), “Variable Selection via Gibbs Sampling”, *Journal of the American Statistical Association*, 88, 881-889.

Geweke, J. (1988), “Exact Inference in Models with Autoregressive Conditional Heteroscedasticity”, in. E. Berndt, H. White, W. Barnett (eds.), *Dynamic Econometric Modeling,* Cambridge University Press, Cambridge.

Hausman, J.A. (1978), “Specification Tests in Econometrics”, *Econometrica*, 46, 1251-1272.

Hodges, J.S. (1990), “Can/May Bayesians Do Pure Tests of Significance?”, in S. Geisser, J.S. Hodges, S.J. Press and A. Zellner (eds.), *Bayesian and Likelihood Methods in Statistics and Econometrics*, North-Holland, Amsterdam, 75-90.

Houston, F.S. and D.L. Weiss (1975), “Cumulative Advertising Effects: The Role of

Serial Correlation", *Decision Sciences*, 6, 471-481.

Inder, B.A. (1990) "A New Test for Autocorrelation in the Disturbances of the Dynamic Linear Regression Model", *International Economic Review*, 31, 341-354.

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

Kiefer, N.M. and J.F. Richard (1987), "Decision Theory, Estimation Strategies and Model Choice", CAE Working Paper 87-08, Cornell University, Ithaca, NY.

King, M.L. (1983), "Testing for Autoregressive Against Moving Average Errors in the Linear Regression Model", *Journal of Econometrics*, 21, 35-51.

King, M.L. (1987-1988), "Towards a Theory of Point Optimal Testing", *Econometric Reviews*, 6, 169-255 (with discussion).

Koop, G., J. Osiewalski and M.F.J. Steel (1994), "Bayesian Long-Run Prediction in Time Series Models", *Journal of Econometrics*, forthcoming.

Lempers, F.B. (1971), *Posterior Probabilities of Alternative Linear Models*, Rotterdam University Press, Rotterdam.

Min, C. and A. Zellner (1993), "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates", *Journal of Econometrics*, 56, 89-118.

Mitchell, T.J. and J.J. Beauchamp (1988), "Bayesian Variable Selection in Linear Regression, *Journal of the American Statistical Association*, 83, 1023-1036 (with discussion).

Monahan, J.F. (1983), "Fully Bayesian Analysis of Time Series Models", *Journal of Econometrics*, 21, 307-331.

Osiewalski, J. and M.F.J. Steel (1992), "A Bayesian Note on Competing Correlation Structures in the Dynamic Linear Regression Model", *Economics Letters*, 40, 383-388.

Osiewalski, J. and M.F.J. Steel (1993a), “Regression Models Under Competing Covariance Structures: A Bayesian Perspective”, *Annales d’Economic et de Statistique*, 32, 65-79.

Osiewalski, J. and M.F.J. Steel (1993b), „Robust Bayesian Inference in Elliptical Regression Models”, *Journal of Econometrics*, 57, 345-363.

Palda, K.S. (1964), *The Measurement of Cumulative Advertising Effects*, Prentice Hall, Englewood Cliffs.

Palm, F.C. and A. Zellner (1992), “To Combine or Not to Combine? Issues of Combining Forecasts”, *Journal of Forecasting*, 11, 687-701.

Pericchi, L.R. (1984), “An Alternative to the Standard Bayesian Procedure for Discrimination Between Normal Linear Models”, *Biometrika*, 71, 575-586.

Pettit, L.J. (1992), “Bayes Factor for Outlier Models Using the Device of Imaginary Observations”, *Journal of the American Statistical Association*, 87, 541-545.

Poirier, D.J. (1985), “Bayesian Hypothesis Testing with Consistent Priors Across Models”, in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (eds.), *Bayesian Statistics 2*, North-Holland, Amsterdam, 711-722.

Poirier, D.J. (1988a), “Bayesian Diagnostic Testing in the General Linear Normal Regression Model”, in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (eds.), *Bayesian Statistics 3*, Clarendon Press, Oxford, 725-732.

Poirier, D.J. (1988b), “Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics”, *Journal of Economic Perspectives*, 2, 121-170 (with discussion).

Poirier, D.J. (1992), “Window Washing: A Bayesian Perspective on Diagnostic Checking”, mimeo.

Richard, J.F. (1979), “Exogeneity, Inference and Prediction in So-Called Incomplete

Dynamic Simultaneous Equation Models", CORE Discussion Paper 7922, Université Catholique de Louvain, Louvain-la-Neuve.

Spiegelhalter, D.J. and A.F.M. Smith (1982), "Bayes Factors for Linear and Log-Linear Models with Vague Prior Information", *Journal of the Royal Statistical Society*, B 44, 377-387.

Tierney, L. and J.B. Kadane (1986), "Accurate Approximations for Posterior Moments and Marginal Densities", *Journal of the American Statistical Association*, 81, 82-86.

van den Broeck, J., G. Koop. J. Osiewalski and M.F.J. Steel (1994), "Stochastic Frontier Models: A Bayesian Perspective", *Journal of Econometrics*, 61, forthcoming.

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

Zellner, A. (1984), *Basic Issues in Econometrics*, University of Chicago Press, Chicago.

Zellner, A., C. Hong and G.M. Gulati (1990), "Turning Points in Economic Time Series, Loss Structures and Bayesian Forecasting", in S. Geisser, J.S. Hodges, S.J. Press, A. Zellner (eds.), *Bayesian and Likelihood Methods in Statistics and Econometrics*, North-Holland, Amsterdam, 371-393.

Zellner, A. and A. Siow (1980), "Posterior Odds Ratios for Selected Regression Hypotheses", in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (eds.), *Bayesian Statistics*, Valencia University Press, Valencia, 585-603.

FIGURE 1: POSTERIOR DENSITY OF ETA

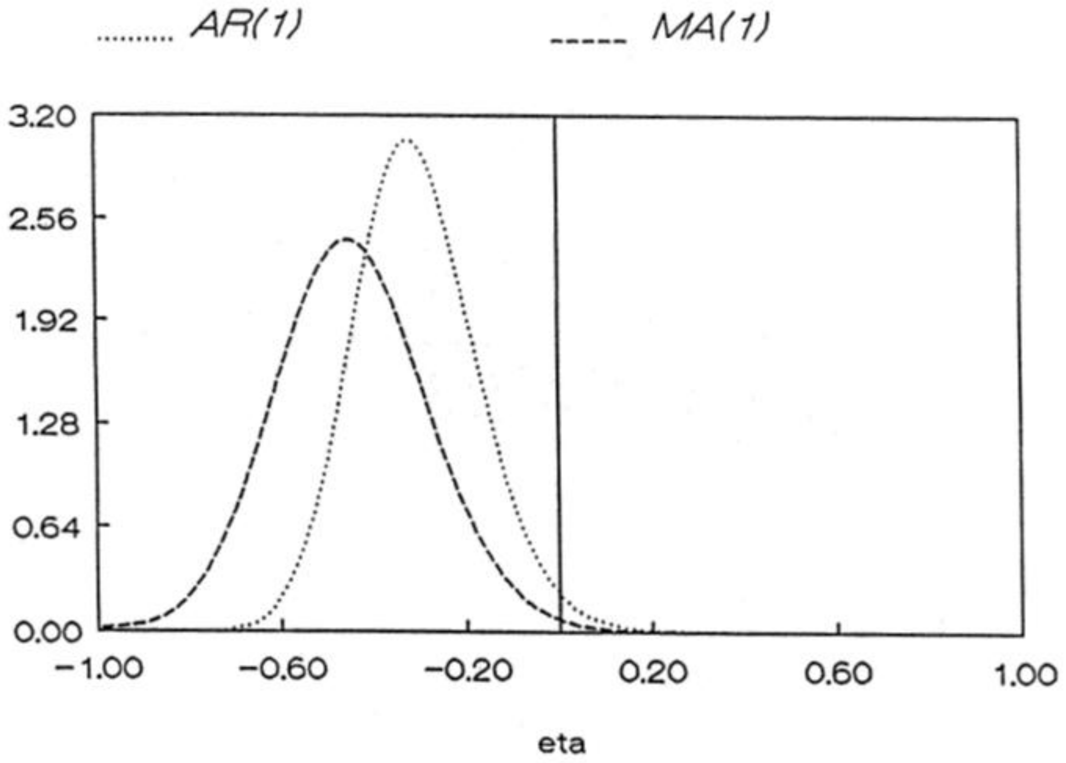
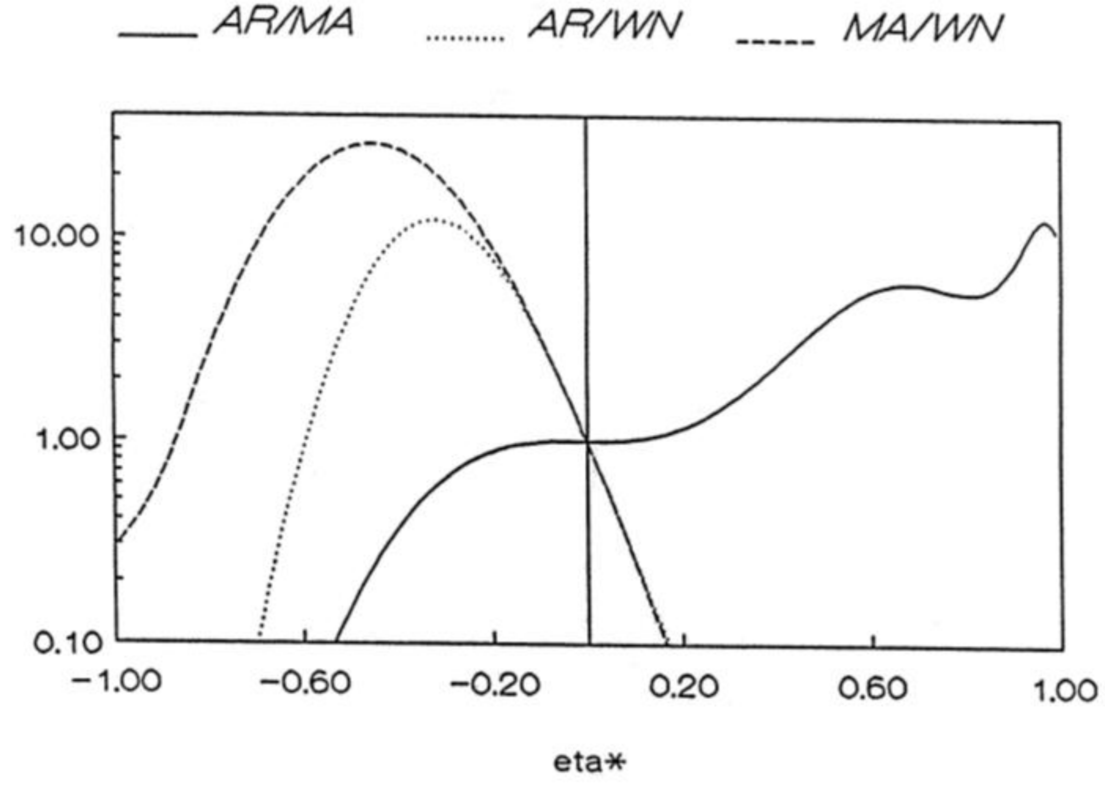


FIGURE 2: CONDITIONAL BAYES FACTOR
log scale


AR/MA
AR/WN
MA/WN
eta*