Bayesian Testing and Testing Bayesians†

by

Jean-Pierre Florens*
and
Michel Mouchart**

April 1992

ABSTRACT

This paper surveys Bayesian procedures of testing along with different attitudes of econometricians facing testing problems with some sympathy for Bayesian ideas. In the first part, the general Bayesian testing procedures are structured along two main axes. The first one presents the usual procedures based on the posterior probabilities of models, or of hypothesis, and relying mainly on a comparison of predictive distributions relative to each model, or hypothesis. The second one has been developped more recently and is based on a Bayesian extension of the encompassing principle. This approach basically relies on a comparison of posterior distributions on parameters of interest conditional on each hypothesis.

In the second part, different attitudes of testing Bayesians is considered for the testing of two important problems in econometrics. The first one, testing the significance of a regression coefficient and chosing the regressors, is used to exemplify the flexibility of the Bayesian approach whereas the second one, testing for unit root, is used to exemplify how Bayesian thinking may be useful to shed new insights on controversial issues. A final section provides offers a short survey of Bayesian contributions on other testing problems.

This text presents research results of the Belgian programme on interuniversity poles of attraction initiated by the Belgian state, Prime Minister's Office, Science Policy Programming. The scientific responsability is assumed by its authors.

---

† The authors are particularly indebted to J.H Drèze for discussions and comments on a preliminary version and to L. Bauwens, S. Larribeau-Nori, E. Moreno and J. Rodriguez-Poo for some very useful remarks. The authors are also deeply indebted to J.F. Richard with whom many ideas involving the encompassing principles have been elaborated. They are also thankful to J. Bouoiyour for bibliographical assistance. The authors retain however full responsibility, in particular for the kind of presentation they have finally adopted.

* GREMAQ, Université des Sciences Sociales, Toulouse, France

** CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgique and GREMAQ, Université des Sciences Sociales, Toulouse, France.

ISSN-0771 3894

1. Introduction

In general, a null hypothesis (H_0) and its alternative (H_1) may be considered as two region (actually, a partition) of the parameter space of a unique sampling model or as two different models. These two interpretations cover a large class of problems and a wealth of different procedures have been suggested and analyzed in sampling theory. This survey focuses the attention on Bayesian methods and endeavours to illustrate both the intrinsic unity of Bayesian thinking and its basic flexibility to adjust to and to cope with a wide range of circumstances.

Broadly speaking two kinds of approaches seem to be fruitful. One approach starts by remarking that a test is a statistical procedure -*i.e.* a function defined on a sample space- with value in a two points space the elements of which may be, more or less arbitrarily, labelled "accept" or "reject" a given hypothesis. It seems therefore natural to approach a testing problem as a two decisions problem : this approach gives a rather straightforward logical background and leads to what we shall call a "fully Bayesian approach". Some colleagues among our Bayesian friends are tempted to assert "that's all" and to consider testing exclusively as a two decisions problem. Another approach starts by remarking that testing may also be viewed as a way for the statistician to handle his own doubt on a statistical model : were the null hypothesis considered as certain, then he would go on with an inference restricted to the null hypothesis *i.e.* with a statistical model the parameters of which sweep the null hypothesis only. Thus the statistician implicitly considers two models ;

$$\left\{P^\theta : \theta \in H_0\right\} \text{ and } \left\{P^\theta : \theta \in H_1\right\}$$

and the mere fact of testing means that (s)he is uncertain about which model is the "suitable" one. We shall see that this approach is, in a sense, more flexible. It is characterized by the fact that the two models share a same sample space and two parameter spaces characterizing different sets of sampling distributions but these parameters may possibly have an economic meaning different with the model even if they look like as if they were partially common to the two models. As suggested by the title, this survey is made of two main sections : one on general principles, another one on procedures. Let us briefly present those sections.

Section 2, called Bayesian testing, surveys general principles and is divided into two main trends. The first one (section 2.2) is rather traditional in the Bayesian literature and focuses the attention directly on the inference on the hypotheses in presence. Its main output is the posterior probability of hypotheses but may also

take the form of Bayes factors or of posterior odds and is based, essentially, on the difference between the predictive distributions determined by each hypotheses. Side issues also consider comparison with sampling theory analogue, in particular with possible conflict with p-values. A second trend (section 2.3), less traditional, focuses the attention on the inference on the parameter of interest under different models and is based on a Bayesian approach to the encompassing principle. This is a formalization of a simple idea : as far as inference is concerned, model M_0 encompasses model M_1 if any inference based on M_1, *i.e.* any of its posterior distributions, may be obtained by transforming a posterior distribution of model M_0 without retrieving the sample, in other words, if M_0 is a "new" model it should be able, for being accepted, to explain the findings of the "old" model M_1. A peculiarity of the encompassing approach is to base the binary choice among models, or hypotheses, on its relevance for the inference on the parameters of interest. These two trends are distinguished by writing (H_0, H_1) for the first one and (M_0, M_1) for the second one but it should be insisted that this notational distinction is contextual rather then substantial, *i.e.* it is aimed at drawing the attention on the question whether the interest lies in choosing among models, understood as hypotheses to "explain the real world" or in choosing among models understood as learning rules to be used when analyzing data. Also section 2 discusses the "minimal amount" of prior specification required to make these approaches operational and some aspect of their asymptotic behaviour.

Section 3, called Testing Bayesians, surveys procedures derived from the general principles of section 2 when facing specific problems of econometric interest. Section 3.1 considers the testing of a regression coefficient (with known variance). This problem is analytically the simplest one ; it is used as an opportunity to survey most of the procedures suggested by the preceding section. In contrast, section 3.2 considers an analytically more complex problem, namely testing for unit root(s) ; it surveys a literature that recently has been both hot and abundant. It is therefore too early to claim any "objectivity" in the appraisal of those recent achievements. We essentially endeavour to show how the general principles of section 2 have generated useful and workable analysis of a problem, the solution of which has been judged of crucial importance in contemporary dynamic modelling.

It should be clear, from this presentation, that section 2 and 3 do not make full justice of the complete Bayesian activity in the domain of testing. The last section sketches some of the overlooked material and proposes a rapid evaluation of the present state-of-the-art.

2. Bayesian Testing

2.1 Introduction

In sampling theory, a *statistical model* may be viewed as a family of probability measures on a sample space $(S, \mathcal{S})$ indexed by a parameter θ :

$$\big\{(S, \mathcal{S}), \mathrm{P}^{\theta} : \theta \in \Theta\big\}. \tag{2.1}$$

In usual cases, these probability measures may be represented by densities on a random variable x representing the observations on the sample space :

$$\big\{p(x \mid \theta) : \theta \in \Theta\big\}. \tag{2.2}$$

Most *econometric models* are of the conditional type, *i.e.* observation x is decomposed into y and z, thus : $x = (y, z)$, and the statistical model is only partially specified, *i.e.* up to the conditional distribution of y given z only, thus :

$$\big\{p(y \mid z, \theta) \;\; \theta \in \Theta\big\}. \tag{2.3}$$

Bayesian models are obtained by endowing a sampling theory model with a (prior) probability measure P_{θ} on the parameter space $(\Theta, \mathcal{T})$ and interpreting the sampling model as a probability conditional on the parameter. This is allowed by defining a unique probability measure Π on the product space $\Theta \times S$, which is written as :

$$\Pi = P_{\theta} \otimes P^{\theta} \tag{2.4}$$

and characterized by :

$$\Pi(A \times B) = \int_{\theta \in A} P^{\theta}(B) \quad dP_{\theta} \qquad A \in \mathcal{T}, \qquad \mathcal{B} \in \mathcal{S} \tag{2.5}$$

or, in terms of densities :

$$p(x, \theta) = p(\theta) p(x \mid \theta). \tag{2.6}$$

Thus, in econometrics, a typical bayesian model becomes a unique joint probability on y and θ conditional on z, *i.e.* in terms of densities :

$$p(y, \theta \mid z) = p(\theta \mid z) \;\; p(y \mid z, \theta). \tag{2.7}$$

Bayesian methods aim at performing a dual decomposition of this joint distribution into a predictive distribution on $(y \mid z)$ and posterior distributions on $(\theta \mid y, z)$:

$$p(y, \theta \mid z) = p(y \mid z)p(\theta \mid y, z). \tag{2.8}$$

Inference problems are connected with the analysis of the transformation "prior to posterior" *i.e.* the transformation of $p(\theta \mid z)$ into $p(\theta \mid y, z)$. Prediction problems may be considered either in prior terms, *i.e.* predicting y before it has been observed, or in posterior terms, *i.e.* decomposing y into a "past" component, y_p, and a "future" component, y_f, $y = (y_f, y_p)$, and predicting y_f after observing y_p. In simple cases, predictions are considered conditionally on z and are handled by averaging the sampling distribution $p(y \mid z, \theta)$ - or $p(y_f \mid y_p, z, \theta)$ - weighted by the relevant distributions on the parameters $p(\theta \mid z)$ - or $p(\theta \mid y_p, z, \theta)-$ in order to obtain the predictive distributions :

$$p(y \mid z) = \int p(y \mid z, \theta)\ p(z \mid \theta) \quad d\theta \tag{2.9}$$

$$p(y_f \mid y_p, z) = \int p(y_f \mid y_p, z, \theta)\ p(\theta \mid y_p, z) \quad d\theta \tag{2.10}$$

A rather precise exposition of the probabilistic foundation of Bayesian methods may be found in Florens, Mouchart and Rolin (1990), particularly chapter 1, to be referred to as EBS. An exposition specifically oriented towards econometrics may be found in Zellner (1971) and with a more general orientation, in Berger (1985), de Groot (1970) Jeffreys (1967) or Lindley (1971).

Remark 1. It will be seen later that both the posterior distributions *and* the predictive distributions are involved in Bayesian procedures of testing. ■

Remark 2. Up to now we have not assumed the prior independence of θ and z, *i.e.* $\theta \perp\!\!\!\perp z$ or $p(\theta \mid z) = p(\theta)$. This is motivated by two types of considerations. First, at this level of generality, such an assumption does not simplify substantially the Bayesian model because even when θ and z are, a priori, independent, all other distributions involved in the Bayesian model, namely the sampling, the predictive and the posterior, still depend crucially on z ; in particular, for the inference *i.e.* the evaluation of the posterior distribution, y *and* z are taken as data. Furthermore in sequential modelling, z would involve lagged values of y so that θ would not be independent of z. Second, the rationale of the hypothesis $\theta \perp\!\!\!\perp z$ is to ensure the admissibility of conditioning on z, *i.e.* of considering the joint distribution $p(y, \theta \mid z)$ rather than the joint distribution $p(y, z, \theta, \lambda)$ where λ would characterize the marginal sampling distribution of z, with

the consequence that z becomes exogenous for the inference on θ. As the problem of exogeneity is not a basic issue in this survey, there is no real gain in assuming the exogeneity of z as a starting point (for a Bayesian treatment of exogeneity see EBS sections 3.4. and 6.3.). As a consequence, most general ideas around testing may be presented in unconditional as well as in conditional models. ∎

2.2 Posterior Probability of Hypotheses, Posterior Odds, Bayes Factors.

2.2.1 The traditional Bayesian approach: comparison of predictive distributions

Two decisions problems

Consider a two-decisions problem with the set of possible decisions $A = \{a_0, a_1\}$. The loss function may then be written as $l(a_j, \theta) = \tilde{l}_j(\theta)$ for $j = 0, 1$ and H_0 and H_1 may be *defined* as $H_j = \tilde{l}_j^{-1}(\{0\})$ *i.e.* the set of θ for which a_j is optimal (*viz.* has zero loss). A testing problem is characterized by the fact that H_0 and H_1 operate a partition of the parameter space constituting the maintained hypothesis. Thus a_j may be *interpreted* as "decide *as if* H_j were true" and the loss function may be rewritten as :

$$l(a_j, \theta) = \ell_j(\theta) \mathbb{1}_{H_{\bar{j}}}(\theta) \tag{2.11}$$

where $\bar{j} = 0$ if $j = 1$ and $\bar{j} = 1$ if $j = 0$. A decision function a: $S \to A$ is an optimal one, a^*, when defined as :

$$a^*(x) = arg \inf_{a \in A} E[l(a, \theta) \mid x] \qquad \text{, in general.} \tag{2.12}$$

i.e., in the case of a two-decisions problem :

$$a^*(x) = a_1 \iff E[l_1(\theta) \mathbb{1}_{H_0}(\theta) \mid x] \leq E[l_0(\theta) \mathbb{1}_{H_1}(\theta) \mid x]. \tag{2.13}$$

It has been argued that Neyman-Pearson theory may be interpreted in this decision theoretic approach by considering the particular case of a piece-wise constant loss function $l(a_j, (\theta)) = l_j \mathbb{1}_{H_{\bar{j}}}(\theta)$ so that

$$a^*(x) = a_1 \Longleftrightarrow l_1 P(H_0 \mid x) \leq l_0 P(H_1 \mid x) \ , \tag{2.14}$$

i.e. in terms of *posterior probability*:

$$a^*(x) = a_1 \Longleftrightarrow P(H_0 \mid x) \leq \frac{l_0}{l_0 + l_1} \ , \tag{2.15}$$

or in terms of *posterior odds*:

$$a^*(x) = a_1 \Longleftrightarrow \frac{P(H_0 \mid x)}{P(H_1 \mid x)} \leq \frac{l_0}{l_1} \ . \tag{2.16}$$

Thus the rule $a^*(x) = a_1 \Longleftrightarrow P(H_0 \mid x) \leq 5\%$" would be justified with the piece-wise constant loss function such that $l_1 = 19l_0$, a way of formalizing that type I-error is much more severe than type II-error. This analysis shows that Bayesian testing under piece-wise constant loss functions boils down to the evaluation of the posterior probability of the hypotheses in presence.

Posterior probabilities and posterior odds as foundations of Bayesian testing are as old as Bayesian statistics and are presented e.g. in Jeffreys (1967). These definitions are given and discussed in many books and used in many papers : see *e.g.* Berger (1985), Hodges (1990), Leamer (1978,1983,b and 1991), Zellner (1984). Asymptotic properties of posterior probabilities of the two hypotheses can be found in EBS section 7.5. Note that the posterior odds essentially mimic the likelihood ratio with however a crucial difference of interpretation : the posterior odds are based on the posterior distribution rather than on the sampling distribution and the critical value is defined by the loss structure rather that by an exogenously given level. Also should be stressed the particular piece-wise constant structure of the underlying loss function, an often disputable structure. For example, such a structure would be questionable when testing $H_0 : \theta = 0$ for θ being a price elasticity in a brand demand equation but would be rather reasonable for testing $H_0 : \theta \leq .5$ when θ is the percentage of favourable intention of votes in a forthcoming two-candidates election and when the underlying decision is a possible reorientation of the electoral campaign. The connection between testing and decision is discussed in particular in Savage (1954), Rubin (1971).

Prior specification and structure of the hypotheses

The literature on the posterior probability of hypotheses and on posterior odds has also paid some attention to the role of the prior specification, particularly for the case of a sharp null hypothesis and for improper prior distribution. (see *e.g.*, Klein and Brown (1984), Leamer (1978, 1983,b), Maddala (1976)).

Once it is recognized that the hypotheses H_0 and H_1 operate a partition of the parameter space Θ, the prior distribution (and also the posterior) may be decomposed conformably with that partition, namely :

$$P_\theta = \pi_0 P_\theta^0 + (1 - \pi_0) P_\theta^1 \tag{2.17}$$

where $\pi_0 = P_\theta(H_0)$ and $P_\theta^j(A) = P(\theta \in A \mid H_j)$, or in terms of density :

$$p(\theta) = \pi_0 \; p^0(\theta) \; \mathbb{1}_{H_0}(\theta) \; + \; (1-\pi_0) \; p^1(\theta) \; \mathbb{1}_{H_1}(\theta) \tag{2.18}$$

where $p^j(\theta)\mathbb{1}(\theta)$ are the respective conditional densities. One then obtains :

$$P(H_0 \mid x) = \frac{\pi_0 \; p(x \mid H_0)}{\pi_0 \; p(x \mid H_0) + (1-\pi_0) \; p(x \mid H_1)} \tag{2.19}$$

or equivalently, in terms of posterior odds

$$\begin{aligned} \frac{P(H_0 \mid x)}{P(H_1 \mid x)} &= \frac{\pi_0}{1-\pi_0} \cdot \frac{p(x \mid H_0)}{p(x \mid H_1)} \\ &= \frac{\pi_0}{1-\pi_0} B_x(H_0; H_1) \end{aligned} \tag{2.20}$$

where

$$p(x \mid H_j) = \int_{\theta \in H_j} p(x \mid (\theta) p^j(\theta) \mathbb{1}_{H_j}(\theta) d\theta \qquad j = 0, 1 \; . \tag{2.21}$$

are the predictive densities relative to each hypothesis and $B_x(H_0; H_1)$, called the *Bayes factor* of H_0 against H_1, is deemed to measure how far an observation x increases or decreases the prior odds $\pi_0(1-\pi_0)^{-1}$ into the posterior odds, *i.e.* gives empirical evidence more or less in favour of H_0. Thus in case of piece-wise constant loss functions, the Bayes solution, in terms of posterior probability (2.15), posterior odds (2.16) or Bayes factor (2.20), essentially relies on the predictive densities (2.21) of each models. Let us now survey the impact of the prior specification on those predictive densities according to the structure of the hypotheses to be tested.

(i) If both hypotheses H_j are simple -*i.e.* $H_j = \{\theta_j\}$ so that $\Theta = \{\theta_0, \theta_1\}$- the Bayes factor is exactly equal to the likelihood ratio and is therefore independent of any prior specification. In this case, the Bayes factor and the likelihood ratio give the same empirical evidence in favour of H_0 but again the critical value for that likelihood ratio depends on the loss structure *and* on the prior weight π_0, *viz*

$$a^*(x) = a_1 \Longleftrightarrow B_x(H_0; H_1) \equiv \frac{p(x \mid \theta_0)}{p(x \mid \theta_1)} \leq \frac{l_0}{l_1} \cdot \frac{1-\pi_0}{\pi_0} \; . \tag{2.22}$$

(ii) When one of the hypotheses (or both) is not simple, the corresponding predictive density (2.21) clearly depends on the prior specification. If P_θ^j is improper, *i.e.*

$$\int p^j(\theta) \mathbb{1}_{H_j}(\theta) \quad d\theta = \infty \; , \tag{2.23}$$

more precisely is a σ-finite measure rather than a probability measure, the decomposition (2.17) cannot anymore be interpreted as a marginal-conditional decomposition with respect to H_0 and H_1 ; and $p(x \mid H_j)$ is also improper. But once an improper prior distribution is justified by an invariance argument, its density is defined up to a multiplicative constant only ; for instance when $\Theta = I\!R$, $p^j(\theta) = 1$ corresponds to Lebesgue measure and has the same (translation) invariance property as $p^j(\theta) = 2$ (or, =1991). Furthermore, two improper prior distributions with proportional densities are associated with identical posterior distributions and therefore a Bayesian cannot argue in favour of $p^j(\theta) = 2$ (or, =1991). This is actually reflected by writing $p^j(\theta) \propto 1$ instead of $p^j(\theta) = 1$. The difficulty is that the (2.21) suffers from the same indeterminacy and the Bayes factor is eventually indeterminate.

(iii) Consider now the case where H_0 is sharp but H_1 is not (heuristically the Lebesgue measure of H_0 is zero and that of H_1 is strictly positive ; for instance : $\Theta = I\!R$ and $H_0 = \{\theta_0\}$). If P_θ is smooth (more precisely, dominated by Lebesgue measure) $\pi_0 = P(H_0) = 0$; this implies, with smooth sampling probabilities, that $P(H_0 \mid x) = 0$ and that the posterior odds and the Bayes factor are identically zero. Conversely, if $\pi_0 = P(H_0) > 0$, P_θ is not smooth but $\pi_0 > 0$ whereas P^1_θ is smooth implies a discontinuity of P_θ which has been criticized in some contexts. Suppose, for instance, $\Theta = I\!R$, $H_0 = \{\theta_0\}$, $\pi_0 = .5$ and $(\theta \mid H_1) \sim N(m, v)$. The prior distribution function is smooth everywhere except at the point θ_0 where there is a jump of height equal to .5. This jump may be considered as undesirable if the sampling model is smooth, heuristically if the likelihood function is continuous at $\theta = \theta_0$. For instance, when equilibrium appears as a parametric restriction of the type $g(\theta) = 0$ in a model embodying structural disequilibrium, a prior distribution with a strictly positive mass on the (null measure) surface $g(\theta) = 0$ may be palatable whereas in a demand model where θ_i represents a price elasticity, a strictly positive prior probability on $\{\theta_i = -1\}$ may be much more questionable when the prior distribution is otherwise continuous. This suggests that the specification of the point of discontinuity of the prior distribution would be associated to the question whether one is willing to test $H_0 \ : \ \theta_i = -1 \quad H_1 \ : \ \theta_i \neq -1$ or $H_0 \ : \ \theta_i = 0 \quad H_1 \ : \ \theta_i \neq 0$.

Alternative approaches to testing

It has been suggested that testing a sharp hypothesis might be a workable approximation for testing a neighborhood *i.e.* $H_0 \ : g(\theta) = 0$ would be an approximation for $H_0 :\mid g(\theta) \mid < \varepsilon$ for some $\varepsilon > 0$ so that π_0 should be interpreted as $P[-\varepsilon < g(\theta) < \varepsilon] = \pi_0$ rather than $P(g(\theta) = 0) = \pi_0$. The difficulty is that

the optimal decision rule $a^*(x)$, the posterior probability, posterior odds or Bayes factors all depend crucially on ε. Thus for the case of sharp hypothesis some statisticians have remarked that in a sampling theory approach a yes-no testing at a predefined critical level α may be less interesting than the evaluation of a confidence interval of given level. Similarly, in a Bayesian framework a yes-no decision problem may be less interesting, in practice, than the evaluation of a posterior probablity interval *i.e.* the evaluation of $c_1(x)$ and $c_2(x)$ (with $c_1(x) < c_2(x)$) such that $P[c_1(x) \leq g(\theta) \leq c_2(x) \mid x] = 1 - \alpha$. (see *e.g.* Lindley 1961). In this approach, the choice of level α is less crucial: it is introduced in the framework of a convenient summary of the complete posterior distribution.

Instead of confining the attention to the two-decisions problem , attention has also been paid to the more cognitive problem of the empirical evidence provided by the data in favour of H_0 or of H_1. Comparing sampling theory and Bayesian approaches suggests to compare the p-value, considered as a measure of empirical evidence, and one of the Bayesian natural measures *viz* posterior probability of H_0, posterior odds or Bayes factor. A vast litterature has shown that, on many respects, an observation may look like very favourable to the alternative hypothesis in a sampling theory framework but very favourable to the null hypothesis in a Bayesian framework ; for given data a small p-value may be associated to a very high Bayes factor, so that a p-value cannot be viewed as an approximation of $P(H_0 \mid x)$, not even an indication of its order of magnitude. This situation has been known as Lindley's paradox, after Lindley (1957) and has received an impressive attention in the literature. Fundamentally, the problem is that any posterior distribution may be associated to a given likelihood function.

Relationships between Bayesian tests and classical tests have been discussed in many papers, see in particular Bernardo (1980), de Groot (1973), Gaver and Geisel (1974), Good (1985), Hill (1982, 1990), Poirier (1988) ; for the more particular problem of the relationship between p-value and Bayes factors, see Berger (1986), Berger and Berry (1988) Berger and Delampady (1987) (this paper also contains an important bibliography on this issue), Berger and Sellke (1987), Casella and Berger (1987) Dickey (1977), Shafer (1982). Moreno and Cano (1986) have however shown that multidimensional parameters might reverse the relationship previously reported.

Another approach takes the point of view of estimating of a binary valued function of the parameter $\beta = b(\theta)$ where $b : \Theta \to \{0, 1\}$ is such that $H_j = b^{-1}(\overline{j}) \quad j = 0, 1$, *i.e.* $b(\theta) = \mathbb{1}_{H_0}(\theta)$. Thus some Bayesians have suggested to enlarge the two-points

set of possible decisions into the unit interval $[0,1]$ and to consider loss functions of the form $\ell(a, \mathbb{1}_{H_0}(\theta))$ with $a \in [0,1]$. Thus for a quadratic loss function, $\ell(a, \mathbb{1}_{H_0}(\theta)) = [a - \mathbb{1}_{H_0}(\theta)]^2$, the Bayesian solution, is simply $a^*(x) = E(\mathbb{1}_{H_0}(\theta) \mid x) = P(H_0 \mid x)$ and for

$$\begin{aligned} \ell(a, \mathbb{1}_{H_0}(\theta)) &= \ell_0(a - \mathbb{1}_{H_0}(\theta)) \quad \text{for } a > \mathbb{1}_{H_0}(\theta) \\ &= \ell_1(\mathbb{1}_{H_0}(\theta) - a) \quad \text{for } a < \mathbb{1}_{H_0}(\theta), \end{aligned}$$

the Bayesian solution is a corner solution, namely:

$$\begin{aligned} a^*(x) = 0 \text{ if } P(H_0 \mid x) &\leq \frac{\ell_0}{\ell_0 + \ell_1} \\ = 1 \text{ otherwise} \quad . \end{aligned}$$

This is exactly (2.15) with $a = 0$ (resp. $a = 1$) identified with a_1 (resp. a_0). For more details, see Hwang, Robert, Wells and Farrel (1992).

2.3. Inference on genuine parameters of interest : an encompassing approach.

2.3.1. Model choice and hypothesis testing.

A deep and recurrent theme of discussions in econometric modelling turns around the relationship between the economic meaning of a parameter and the model where it appears. These discussions suggest that parameters may possibly change their meaning under a null hypothesis. For instance when testing $\gamma = 0$ in the model $y = \beta z_1 + \gamma z_2 + \varepsilon$ the coefficient of z_1 may have a different meaning according to whether $\gamma = 0$ or $\gamma \neq 0$.

These considerations lead to analyze statistical models such that the "parameter" θ has the form $(\lambda, \theta_0, \theta_1)$ where λ is a "model label" ($\lambda \in \{0,1\}$) and $\theta_j (j = 0,1)$ are model-specific parameters. Thus the full Bayesian model would be specified through a joint distribution $p(\lambda, \theta_0, \theta_1, x)$, written as :

$$p(\lambda, \theta_0, \theta_1, x) = \left[\pi_0 \; p^0(\theta_0, \theta_1, x)\right]^{1-\lambda} \left[(1 - \pi_0) \; p^1(\theta_0, \theta_1, x)\right]^{\lambda} , \tag{2.24}$$

where $\pi_0 = P(\lambda = 0)$ and $p^j(\theta_0, \theta_1, x) = p(\theta_0, \theta_1, x \mid \lambda = j)$ $j = 0,1$, with the following characteristics :

$$\begin{aligned} p^j(\theta_0, \theta_1, x) &= p^j(\theta_j) \; p^j(x \mid \theta_j) \; p^j(\underline{\theta}_{\bar{j}} \mid \theta_j) \qquad j = 0,1 \\ &= p^j(x) \; p^j(\theta_j \mid x) \; p^j(\underline{\theta}_{\bar{j}} \mid \theta_j) \; . \end{aligned} \tag{2.25}$$

This is : the sampling distribution of model j, $p^j(x \mid \theta)$, depends only on the corresponding θ_j with the consequence that the conditional distribution $p^j(\theta_{\overline{j}} \mid \theta_j)$ is not revised, in model j, by the observation x. Let us note the following points:

(i) In this formulation θ_0 may be a subvector of θ_1, or a function of θ_1 and represents the free parameters under H_0 in the usual hypothesis testing but this is not a necessity : θ_0 and θ_1 may also represent unrelated parameters characterizing different models.

(ii) When θ_0 is a subvector of θ_1 and when one is willing to specify that θ_0 keeps the same economic meaning whether $\lambda = 0$ or $\lambda = 1$, then the prior specification would be : $p^1(\theta_0 \mid \theta_1)$ is degenerate at θ_0 and $p^0(\theta_1 \mid \theta_0) = p^1(\theta_1 \mid \theta_0)$, the right-hand side being directly derived from $p^1(\theta_1)$.

These features are stressed by writing M_j (for model) instead of H_j (for hypothesis). Recall that this distinction is contextual rather than mathematical but the context of model choice is deemed to be more general, in a sense to be made precise in the sequel . In particular, if the parameter of interest is λ, *i.e.* if the problem is actually deciding whether model M_0 or M_1 is preferred, the discussion of the preceding section applies. More specifically :

$$P(\lambda = 0 \mid x) = \frac{\pi_0 p^0(x)}{\pi_0 p^0(x) + (1-\pi_0) p^1(x)} \ , \tag{2.26}$$

where, again, $p^j(x)$ is the predictive density relative to model j :

$$p^j(x) = \int p^j(x \mid \theta_j) p^j(\theta_j) \ d\theta_j \ . \tag{2.27}$$

Note that for the same reason as in section 2.2 the evaluation of $p^j(x)$ requires $p^j(\theta_j)$ to be a proper (prior) distribution but also that, although formally present in the full Bayesian model, $p^j(\theta_{\overline{j}} \mid \theta_j)$ needs not be specified at all : it is only assumed to be implicitly an (arbitrary) proper distribution.

2.3.2. The encompassing approach : an overview

Let us pursue the approach of two models, M_0 and M_1, characterized by their parameters θ_0 and θ_1 but consider now that the parameter of interest is a function of θ_1, *viz* $\varphi = f(\theta_1)$.

If model specification, prior elicitation, data handling and numerical computation were "free" (or your research budget "unlimited" in time and money) one could apply a "full Bayesian approach", namely : start from the complete joint distribution

$p(\lambda, \theta_0, \theta_1, x)$ sketched in the preceding section, marginalize into $p(\varphi, x)$ and therefrom evaluate $p(\varphi \mid x)$, which can also be decomposed as follows :

$$p(\varphi \mid x) = P(\lambda = 0 \mid x)\ p^0(\varphi \mid x) + P(\lambda = 1 \mid x)\ p^1(\varphi \mid x)\ , \tag{2.28}$$

where $p^j(\varphi \mid x)$ is the posterior distribution of φ obtained from each model separately, *i.e.* $p^j(\theta_0, \theta_1, x)$. Remember from the discussion of section 2.2.1, that the evaluation of $P(\lambda \mid x)$, and therefore of $p(\varphi \mid x)$, is meaningful only under proper prior specification *i.e.* when $p(\lambda, \theta_0, \theta_1)$ is a genuine probability measure. In practical cases, the formidable "if" starting this paragraph makes this solution more often "dully Bayesian" than "fully" Bayesian" !

Apart from the operationality requirement, there is another argument for thinking that the evaluation of $p(\varphi \mid x)$ as in (2.28) is not the end of the story. As far as M_0 and M_1 are two "reasonable" models for explaining how data x has been generated or for helping in predicting a future realization of x, it is a standing trend in the history of science to look for "parsimonious modelling" rather than keeping record of "all reasonable models" and the use of the mixed model (2.28) reflects precisely the idea of keeping track of "all reasonable models" as opposed to a search for as simple a model as possible. (see also Dickey and Kadane (1980), Hill (1985-1986)).

If model 1 were as simple as model 0, inference on φ would then boil down to evaluate its posterior distribution $p^1(\varphi \mid x)$ from the sole model 1 *viz.* from $p^1(\theta_1 \mid x) \propto p^1(\theta_1) p^1(x \mid \theta_1)$; in such a case, neither model 0 nor $p^j(\theta_{\overline{j}} \mid \theta_j)$ would have any role. Suppose now that model 0 is "definitively simpler". A natural question is now : "how far could we build an inference on φ relying on model 0 only ?". The very same question is also relevant when M_0 and M_1 are two intrinsically different models and the statistician is willing to consider M_0 as a "true" (or "valid") model with the idea that if M_0 were "true" and M_1 were "false", the inference on φ should be based on M_0 rather than on M_1, provided that φ can be interpreted under M_0.

Whereas the evaluation, in section 2.3.1, of $P(\lambda \mid x)$, treats symetrically M_0 an M_1, this last question introduces a basic asymetry between M_0 and M_1 : the parameter of interest is a function of θ_1 only and model 0 has *prima facie* an auxiliary role only. At this level of generality θ_0 may be a subvector of θ_1, or a function of θ_1 or a "completely different" parameter and $f(\theta_1)$ might also be a function of θ_0 only : this will become later on the object of particular cases. We consider the Bayesian extension of M_0, as in (2.25) :

$$p^0(\theta_0, \theta_1, x) = p^0(\theta_0)\ p^0(x \mid \theta_0)\ p^0(\theta_1 \mid \theta_0)\ , \tag{2.29}$$

where the first two terms specify a standard Bayesian model on (x, θ_0) and the last term, designated as a *" Bayesian Pseudo True Value"* (BPTV) calls for several remarks :

(i) its name comes from the litterature on the encompassing principle where a pseudo true value of a statistic means its expectation or its probabilistic limit under a model different from that where such a statistic "naturally" appears. It is therefore a function of the parameters of that different model . This concept has known a wide diffusion in the classical literature. Its Bayesian version is in the form of a probability, conditional on the parameters of that different model, rather than a deterministic function of those parameters.

(ii) the role of a BPTV is to insert θ_1 into model 0 in order to make an inference on θ_1 possible from model M_0.

(iii) the role of a BPTV is also to "interpret" θ_1 in the light of model M_0 whereas a classical pseudo true value is rather the interpretation of a statistic under M_0. This role gives full flexibility as far as adjusting the procedure to situation where θ_0 is a function of θ_1 or not, or where θ_1 keeps the same interpretation under M_0 and M_1 or not.

(iv) finally the role of a BPTV is to make θ_0 a sufficient parameter in the Bayesian model extending M_0 (*i.e.* $x \perp\!\!\!\perp \theta_1 \mid \theta_0; p^0$ whereas $x \perp\!\!\!\perp \theta_1 \mid \theta_0; p$ and $x \perp\!\!\!\perp (\theta_1, \lambda); p$ are both false in general).

These ideas have been developped, from a Bayesian point of view in Florens (1988), Florens and Mouchart (1985, 1989), E.B.S. (chap 3), Florens, Mouchart and Scotto (1983), and in various hitherto unpublished papers by Florens, Hendry, Richard and Scotto.

From (2.29) the posterior distribution of φ, $p^0(\varphi \mid x)$ can be evaluated as :

$$
\begin{aligned}
p^0(\varphi \mid x) &= E^0[p^0(\varphi \mid \theta_0) \mid x] \\
&= \int p^0(\theta_0 \mid x) p^0(\varphi \mid \theta_0)\, d\theta_0 \ ,
\end{aligned} \tag{2.30}
$$

where $p^0(\varphi \mid \theta_0)$ is obtained by marginalizing on φ the BPTV $p^0(\theta_1 \mid \theta_0)$. Once we have built two posterior distributions on the parameter of interest, *viz* $p^0(\varphi \mid x)$ and $p^1(\varphi \mid x)$, a natural question consists of asking how different they are.

If they are equal for $\varphi = \theta_1$, this means that any inference possible from model M_1 can be reproduced from model M_0 extended as in (2.29). This also means that

model M_0 extracts from the observation all the information extracted by model M_1 along with the information on θ_0, the genuine parameter of θ_0. In such a case we say that model M_0 *encompasses* model M_1. This concept may be viewed as a statistical translation of the idea that in order to be held as prevailing, a "theory" M_0 should explain not only the observations generated by experiments naturally suggested by itself but also the observations generated by experiments suggested by a "Competing Theory" M_1.

If $p^0(\varphi \mid x)$ and $p^1(\varphi \mid x)$ are not equal, one may measure how different they are by means of a discrepancy function $d(x)$ which may be a distance (D) among distributions or among characteristics of distributions (typically : among expectations), or a divergence function (D) :

$$d(x) = D(p^1(\varphi \mid x), p^0(\varphi \mid x)) \tag{2.31}$$

The basic idea is the following : if $d(x)$ is "small", the inference on φ, produced by model M_1, is "almost" reproduced by model M_0. This suggests another way of testing model M_0 against model M_1, when the parameter of interest is φ and when the statistician is not willing to adopt a fully Bayesian attitude and let the model choice depend on prior probabilities of models.

The use of $d(x)$ for deciding to base the inference about φ on M_0 rather than on M_1 may be guided by the asymptotic behaviour of $d(x)$, and eventually of $p^0(\varphi \mid x)$ and of $p^1(\varphi \mid x)$. Instead of describing a general theory let us sketch the structure of a rather general situation. Let us denote $x_1^n = (x_1 \cdots x_n)$. Under M_0, *i.e.*, under the joint distribution $p^0(\theta_0)p^0(x_1^n \mid \theta_0)$, $p^1(\varphi \mid x_1^n)$ typically (*i.e.* under "good" conditions such as i.i.d. sampling with suitable smoothness of the likelihood) converges to a distribution degenerated at $f[g(\theta_0)]$ where $g(\theta_0)$ is the (classical) pseudo-true value of θ_1, *i.e.* the limit under $p^0(x_1^n \mid \theta_0)$ of the maximum likelihood estimator of θ_1 in $p^1(x_1^n \mid \theta_1)$ (See Berk (1966)). Under similar conditions $p^0(\theta_0 \mid x_1^n)$ also converges to a degenerate distribution, (see, *e.g.*, EBS, section 9.3 for cases more general than the i.i.d. one) but this is not sufficient to ensure the convergence of $p^0(\varphi \mid x_1^n)$ which crucially depends on the specification of the BPTV $p^0(\varphi \mid \theta_0)$ (derived from $p^0(\theta_1 \mid \theta_0)$). This analysis however shows that a"natural" specification of the BPTV $p^0(\theta_1 \mid \theta_0)$ would be either a conditional distribution degenerated at the classical pseudo-true value $g(\theta_0)$ or a sample-size dependent specification $p_n^0(\theta_1 \mid \theta_0)$ such that $p_n^0(\theta_1 \mid x_n^1)$ converges to the same limit as $p^1(\theta_1 \mid x_n^1)$; some examples will be given later on. Under such a specification of the BPTV $p^0(\theta_1 \mid \theta_0)$ - or of $p_n^0(\theta_1 \mid \theta_0)$

- and with a suitable choice of the distance or of the divergence, $d(x_1^n)$ typically converges, under M_0, to a chi-square distribution and diverges to infinity under M_1. This situation is therefore formally similar to most testing problems in a sampling theory approach. It suggests either to evaluate a Bayesian p-value, *i.e.* $P^0(d \geq d(x))$, or to build a critical partition of the sample space $S = S_0^{(k)} \cup S_1^{(k)}$ where

$$S_1^{(k)} = \{x \mid d(x) \geq k\} = d^{-1}([k, \infty)) \tag{2.32}$$

where k is specified in such a way that $P^0(S_1(k))$ obtains a prespecified level α *i.e.* k in such that $P^0(S_1(k)) = \alpha$.

Before illustrating those procedures by some examples, let us point out some remarks :

(i) Those procedures do not require the specification of π_0 because the problem is not any more to choose a model but rather to decide how to make inference on a specific parameter of interest. Even if φ is an injective transformation of θ_1, evaluating the predictive support of an observation x in favour of a model or of an hypothesis is different from evaluating the discrepancy between two inferences.

(ii) As far as the prior specification on (θ_0, θ_1) is concerned, $p^0(\theta_0)$ should be proper because $d(x)$ is to be calibrated against the predictive distribution $p^0(x)$ under M^0. However $p^1(\theta_1)$ might be improper, provided that $p^1(\theta_1 \mid x)$ is well defined (*i.e.* that $p^1(x)$ is finite). As far as the BPTV is concerned, $p^1(\theta_0 \mid \theta_1)$ needs not be specified but $p^0(\theta_1 \mid \theta_0)$ is of crucial importance. Next subsection gives more comments on some issues related to the prior specification.

(iii) As mentionned earlier, a first motivation for model choice is parsimonious modelling. In a Bayesian framework, this leads to the idea of developping procedures based on partially specified Bayesian models. Preceding remark (ii) gives, in this respect, a first rationale for the procedures just sketched. A second motivation for model choice is to introduce the statistician's own doubt on his own model. This is the idea that, in an exploratory study, as opposed to a completely specified decision problem, the statistician may like to entertain the idea of a "small world", within which he is willing to be "coherent" but, at the same time, to keep the idea of a "grand world" where he is actually moving and the description of which is only partially specified in his small world. He is thus implicitly accepting the idea of receiving to-morrow information he would consider as unexpected to-day. Savage (1954, section 2.3 and 5.5) mentions the difficulty of delineating the border between the small world and

the grand world. A possible interpretation of the critical partition $S = S_0(k) \cup S_1(k)$ is that a simple and operational approach of the small world is $S_0(k)$ under M_0 and that an observation in $S_0(k)$ leads to go on learning according to $p^0(\varphi \mid x)$ whereas an observation in $S_1(k)$ would lead to rethink the model without deciding to-day how it should be. In the same line, Pratt, Raiffa and Schlaiffer (1964) mention that the extensive-form analysis requires one more axiom than the normal-form analysis, namely their axiom 5 which says, heuristically, that to-day preferences among conditional lotteries are equivalent to conditional preferences among the corresponding unconditional lotteries. That supplementary axiom deals with the stability of the preferences with respect to the state of information. Again, the critical partition $S = S_0(k) \cup S_1(k)$ may be interpreted as describing the set $S_0(k)$ of observations x where one is willing to retain axiom 5 and eventually to use the learning rule specified by $p^0(\varphi \mid x)$ whereas an observations in $S_1(k)$ would lead to re-think the model without deciding to-day how it should be. Both interpretations makes rather natural an a-priori given level α defining k and its associated partition of the sample space S. The alternative to that specification would be a fully Bayesian approach sketched in the beginning of this subsection but also suggested as being of moderate interest in various actual situations.

2.3.3. Prior specification and Bayesian Pseudo True-Values.

Comparing sampling theory procedures and the Bayesian procedures, one may notice that in both types of procedures, M_0 and M_1 are not treated symetrically. In particular the test statistics is calibrated against M_0 only. The main difference between those two types is the prior specification involved in the Bayesian procedure. The examples, in section 3, illustrate how the prior specification is the main source of flexibility in the working of Bayesian procedures. Two types of prior elicitation are required, *viz*: the prior distributions of each models $p^j(\theta_j) \quad j = 0, 1$ and the Bayesian Pseudo-True value $p^0(\theta_1 \mid \theta_0) \quad$. We shall first discuss the specification of $p^0(\theta_1 \mid \theta_0)$. From a purely subjectivist point of view, $p^0(\theta_1 \mid \theta_0)$ represents, in probabilistic terms, the meaning of θ_1 under M_0, *i.e.* the interpretation of the parameter θ_1 if M_0 were considered as the "true" model. It may nevertheless be helpful to suggest some devices for facilitating that specification. In broad terms, those devices are model-based and ensure reasonable asymptotic properties, *viz.* that $p^0(\varphi \mid x_1^n)$ and $p^1(\varphi \mid x_1^n)$ converge to a same limit. We sketch three such devices.

A first device is based on an idea close to Neyman-Pearson approach, namely that M_0 should be rejected only if there is kind of a compelling reason. Thus one

may look for the BPTV which minimizes the expected discrepancy function $d(x)$:

$$p^0(\theta_1 \mid \theta_0) = arg\ inf\ E^0[d(x)]\ , \tag{2.33}$$

where E^0 means the expectation under $p^0(x)$, and inf runs over the set of all conditional distribution of $(\theta_1 \mid \theta_0)$. This optimization raises, in general, difficult problems in the calculus of variation but an example in section 3.3.1 will show how to obtain a restricted optimum in a specific case.

A second device is based on a Bayesian analogue of a sampling theory concept of pseudo-true value, *i.e.* the (sampling) expectation, under M_0, of the posterior distribution, in M_1, of θ_1 namely :

$$p_n^0(\theta_1 \mid \theta_0) = \int p^1(\theta_1 \mid x_1^n)\ p^0(x_1^n \mid \theta_0)\ dx_1^n\ . \tag{2.34}$$

Section 3.1 gives an example of how operational such a device can be made. It may be mentionned that this construction typically leads to a BPTV depending on the sample size, and on the exogenous variables in case of conditional models, and in "good" cases, produces a sequence of $p^0(\varphi \mid x_1^n)$ converging to a same degenerate limit as the sequence $p^1(\varphi \mid x_1^n)$.

A third class of devices consists of considering degenerate distributions *i.e.*.

$$P^0(\theta_1 \in F \mid \theta_0) = \mathbb{1}_{\{g(\theta_0) \in F\}}\ , \tag{2.35}$$

where $g : \Theta_0 \to \Theta_1$ (Θ_j : parameter space for θ_j) is momentarily , arbitrary. This implies :

$$P^0(\theta_1 \in F \mid x) = P^0[g(\theta_0) \in F \mid x] \quad \text{and}$$
$$P^0(\varphi \in Q \mid x) = P^0[f[g(\theta_0)] \in Q \mid x]\ .$$

A natural specification for the function g is *a* sampling theory pseudo-true value. For instance, $g(\theta_0) = plim^0\ \hat{\theta}_{1,n}$ where $\hat{\theta}_{1,n}$ is a maximum likelihood (or another consistent) estimator of θ_1 in M_1 and $plim^0$ denotes $plim$ under M_0 or, in finite sample, $g_n(\theta_0) = E^0(\hat{\theta}_{1,n} \mid \theta_0)$ or $g_n(\theta_0) = arg\ inf_{\theta_1} D(p^0(x_1^n \mid \theta_0), p^1(x_1^n \mid \theta_1))$ for some divergence or distance function. In "good" cases, all these specifications converge to a same limit and produce a sequence of $p^0(\varphi \mid x_1^n)$ converging to a same degenerate limit as the sequence $p^1(\varphi \mid x_1^n)$.

We like to stress that these three devices are offered as an eventual adjuvant to a purely subjectivist specification : wheter a Bayesian statistician should or should not

take care of the asymptotic properties of his procedure is not an issue to be discussed in this review. We simply suggest that different statisticians facing different real-life situations may be expected to adopt different positions on this issue.

Suppose now that the prior specification of $p^j(\theta_j)$ $(j = 0, 1)$ and $p^0(\theta_1 \mid \theta_0)$ has been completed in some way. Experience with computer-assisted elicitation programs has shown the importance of cross-checking whether a *prima facie* elicitation correctly represents subjective opinions. In this context, it should be realized that the prior specification implies two distributions on $\theta_1 : p^1(\theta_1)$, directly specified, and $p^0(\theta_1) = \int p^0(\theta_0)p^0(\theta_1 \mid \theta_0) \quad d\theta_0$. Thus, an important question is whether these distributions should or should not coãncide (*i.e.* whether $\theta_1 \perp\!\!\!\perp \lambda$ or not). Similarly, when θ_0 is interpreted as a function of θ_1, one should ask whether $p^0(\theta_1 \mid \theta_0)$ and $p^1(\theta_1 \mid \theta_0)$, obtained from $p^1(\theta_1)$, should or should not coãncide (*i.e.* whether $\theta_1 \perp\!\!\!\perp \lambda \mid \theta_0$ or not) and whether $p^1(\theta_0)$ and $p^0(\theta_0)$ should or should not coãncide (*i.e.* whether $\theta_0 \perp\!\!\!\perp \lambda$ or not). These questions are the probabilistic translation of the questions about the meaning of the parameters under alternative models. Finally, the prior specification also involves two predictive distributions, $p^0(x)$ and $p^1(x)$ but note that once they coãncide, inference on λ looses any interests (*i.e.* once $\lambda \perp\!\!\!\perp x$, the prior and posterior probabilities of models coãncide).

3. Testing Bayesians.

3.1. Introduction

In this section we show, through some examples, how testing problems may be handled by statisticians open-minded towards Bayesian ideas. We call these statisticians "Bayesians" and their activity we analyze "testing".

The first example starts from a rather standard problem : testing the significance of a regression coefficient in a multiple regression. Through this seemingly simple problem we try to illustrate the diversity, and propose a survey, of procedures suggested by the general approach sketched in section 2. The next example handles a topic more genuinely relevant in Econometrics, namely the testing for unit root.

3.2. Testing a regression coefficient : an overview of procedures.

3.2.1. A simple problem of testing

Let us consider data $X = (y, Z)$ along with two conditional models :

$$M_0 \ : \ y = bz_1 + u \quad i.e. \ (y \mid Z, \theta_0, M_0) \sim N(bz_1, \sigma^2 I_n) \tag{3.1}$$

$$M_1 \ : \ y = \beta z_1 + \gamma z_2 + u \quad i.e. \ (y \mid Z, \theta_1, M_1) \sim N(\beta z_1 + \gamma z_2, \sigma^2 I_n) \tag{3.2}$$

where : $Z = (z_1, z_2)$ $\theta_0 = b$ $\theta_1 = (\beta, \gamma)$ and with σ^2 known. Model M_0 *may* be interpreted as a restriction of model M_1 by means of "$\gamma = 0$" but this is not always suitable. For instance, an economist (M_1) could suggest that the quantity consumed (y) is explained by the price (z_2) and the income (z_1) and a sociologist could rather suggest that the quantity consumed is essentially a social-class issue and that income is a good proxy for his concept of social-class ; in such a case, the economist and the sociologist have different meaning in mind for the coefficient of z_1. A Bayesian statistician facing these two models may therefore want ot recognize explicitly that these two models are deeply different and that b and β represent different concepts. From an operational point of view, this aspect will be modelled by means of the structure of the prior specification (including the specification of the BPTV). Note that the specification of known (and equal) variance is justified exclusively by an expository motivation : we want an example where all the analytical manipulations boil down to the exploitation of the multivariate normal distribution and nevertheless allow for rather subtle discussions. For the same reason, the prior distributions, in each model, is specified as follows :

$$(b \mid M_0) \sim N(b_p, \sigma^2 h_p^{-1}) \tag{3.3}$$

$$\binom{\beta}{\gamma} \mid M_1 \sim N\Big(\begin{pmatrix} \beta_p \\ \gamma_p \end{pmatrix}, \sigma^2 H_p^{-1} \Big) \ . \tag{3.4}$$

If one wants to specify that the coefficient of z_1 has the same meaning under M_0 and under M_1, one may, although not formally necessary, write β instead of b but in any case, coherence in the prior specification implies that b is distributed under M_0 as $(\beta \mid \gamma = 0)$ under M_1, and consequently that $b_p = \beta_p + h_{12,p}\ h_{11,p}^{-1}\ \gamma_p$ and $h_p = h_{11,p}$ (here $h_{ij,p}$ is the element (i,j) of H_p). In such a case the specification $\gamma_p = 0$ may be recommended for symmetry reasons in which case, (3.3) and (3.4) become :

$$(b \mid M_0) \sim N(\beta_p, \sigma^2\ h_{11,p}^{-1}) \tag{3.5}$$

$$\binom{\beta}{\gamma} \mid M_1 \Big) \sim N\Big[\begin{pmatrix} \beta_p \\ 0 \end{pmatrix}\ ,\ \sigma^2\ H_p^{-1} \Big] \tag{3.6}$$

Note also the following identity :

$$V(\gamma \mid M_1) = h_{11,p}\ \mid H_p \mid_p^{-1} \tag{3.7}$$

In what follows, (3.5) - (3.6) are called, the "coherent prior specification".

3.2.2. Posterior probability of hypotheses.

We first consider the traditional approach by means of posterior probability of hypotheses, posterior odds and Bayes factors. The literature is particularly abundant, see, in particular Dickey (1975), Geisser and Eddy (1979), Lempers (1971), Piccinato(1990) Poirier (1985), Spiegelhalter and Smith (1982), Zellner (1984), Zellner and Siow (1980). Thus we introduce a prior probability on models $\pi_0 = P(M_0)$. In order to evaluate the posterior probability we need the predictive distribution in each models.

$$(y \mid Z, M_0) \sim N[b_p z_1, \sigma^2(I_n + h_p^{-1} z_1 z_1')] \tag{3.8}$$

$$(y \mid Z, M_1) \sim N[\beta_p z_1 + \gamma_p z_2,\ \sigma^2(I_n + Z H_p^{-1} Z')] \ . \tag{3.9}$$

This produces the Bayes factor :

$$B_X(M_0; M_1) = \frac{\mid I_n + h_p^{-1} z_1 z_1' \mid^{-\frac{1}{2}} exp - \frac{1}{2\sigma^2} \parallel y - b_p z_1 \parallel_0^2}{\mid I_n + Z H_p^{-1} Z' \mid^{-\frac{1}{2}} exp - \frac{1}{2\sigma^2} \parallel y - \beta_p z_1 - \gamma_p z_2 \parallel_1^2} \ , \tag{3.10}$$

where

$$\parallel y - b_p z_1 \parallel_0^2 = (y - b_p z_1)(I_n + h_p^{-1} z_1 z_1')^{-1}(y - b_p z_1)$$

$$\parallel y - \beta_0 z_1 - \gamma_p z_2 \parallel_1^2 = (y - \beta_p z_1 - \gamma_p z_2)'(I_n + Z H_p Z')^{-1}(y - \beta_p z_1 - \gamma_p z_2) \ .$$

The Bayes rule (2.13) - (2.16) may then be written as :

$$a^*(X) = a_1 \Leftrightarrow \frac{1}{\sigma^2}\{\parallel y - b_p z_1 \parallel_0^2 - \parallel y - \beta_p z_1 - \gamma_p z_2 \parallel_1^2\} \geq k_n(Z) \tag{3.11}$$

where $k_n(Z) = -2\ ln\big[\frac{l_0}{l_1}\frac{1-\pi_0}{\pi_0}\ \mid I_n + h_p^{-1} z_1 z_1' \mid^{\frac{1}{2}} \mid I_n + Z H_p^{-1} Z' \mid^{-\frac{1}{2}}\big]$.

The l.h.s of (3.11) "looks like" a sampling theory test statistics for testing the hypothesis "$\gamma = 0$" comparing the RSS under M_0 and M_1 with two important differences : (i) here the residuals are taken with respect to the predictive expectation in each model and the weighting matrix is given by the predictive precision matrix (ii) the critical values for the Bayesian decision problem, $k_n(Z)$, blends the utility values l_j, the prior specification π_0, h_p and H_p, and the exogenous variables Z. In particular, for any given data X, variations of π_0, h_p and H_p let $P(M_0 \mid X)$ sweep the whole interval $[0, 1]$.

Let us have a look on the asymptotic behaviour of (3.11). So we denote :

$$V = \frac{1}{n} Z'Z = [v_{ij}] \quad i, j = 1, 2 \ ,$$

$$v_{22.1} = \frac{\mid V \mid}{v_{11}} = v_{22} - v_{12}^2 v_{11}^{-1} \ ,$$

$$M_A = I_n - A(A'A)^{-1}A' \quad A : n \times k \ ,$$

$$c = 2\ ln\frac{\ell_1}{\ell_0} \cdot \frac{\pi_0}{1 - \pi_0} \ .$$

For large n, the Bayesian critical region defined by (3.11) is approximatively described as follows :

$$\frac{1}{\sigma^2} y'(M_{z_1} - M_Z)y \geq c + ln\ h_p - ln \mid H_p \mid + ln\ v_{22.1} + ln\ n\ . \tag{3.12}$$

Using (3.7), under the prior coherent specification (3.5) - (3.6), (3.12) simplifies as follows :

$$\frac{1}{\sigma^2} y'(M_{z_1} - M_Z)y \geq c + ln\ V(\gamma \mid M_1) + ln\ v_{22.1} + ln\ n \tag{3.13}$$

Both in (3.12) and (3.13), the l.h.s. follows a chi-square distribution in the sampling under M_0 so that the probability of rejecting (*i.e.* $a^*(X) = a_1$) tends to zero because of $ln\ n$. Under M_1, the l.h.s. increases as n whereas the r.h.s. increases as $ln\ n$ so that the testing procedure is consistent with a slowly increasing critical value. As mentionned before, a sharp hypothesis is often used in practice as an operational approximation for a "small" interval. In such a case, a frequent criticism against a blind use of classical tests of sharp hypotheses is that for large sample, the null hypothesis is typically rejected, for any test the level of which is independent of the sample size. This is indeed the case for the rather general situation of a consistent test with continuous likelihood function. Thus from a sampling point of view the Bayesian test may also be viewed as an operational way to make the level of the test a (decreasing) function of the sample size while keeping the test consistent. This also explains why, for large samples, a p-value may give quite different a message from a Bayesian test.

This example illustrates some difficulties raised by the use of non-informative prior distributions when evaluating Bayes factor or posterior probabilities of hypothesis. Recall that, in general, the ratio of the predictive densities, under M_0 and M_1, is essentially undefined under non-informative prior specification. Note also that (3.12) is undefined if $h_p = 0$ and/or $\mid H_p \mid = 0$, and that $\ln \frac{h_p}{\mid H_p \mid}$ may have an arbitrary limit if we let h_p and $\mid H_p \mid$ tend jointly to 0. In the case of coherent prior, the critical region, defined by (3.13) tends to become empty once $V(\gamma \mid M_1)$ tends to infinity.

It may be interesting to compare that two-sided test with a one-sided test. A simple case is the following. Let (3.2) (3.4) represent the sampling process and the prior specification respectively. This allows one to evaluate $P(\gamma \geq 0)$ and $P(\gamma \geq 0 \mid X)$ and to construct a Bayesian test with critical region characterized by $\frac{P(\gamma \geq 0 \mid X)}{P(\gamma < 0 \mid X)} < k$ where k is a constant defined as earlier. Note however that, at variance from the two-sided test, this Bayesian test, is unambiguously defined even with improper priors

(such as $P(\beta,\gamma) \propto 1$) and it is easily checked that the posterior probability of the null hypothesis and the sampling p-value are asymptotically equivalent, and that Lindley's paradox eventually vanishes.

3.2.3. Encompassing procedures

General procedures

Suppose first that the parameter of interest is the vector of regression coefficients. Thus in the notation of section 2.4, we retain λ for the model label $\theta_0 = b$, and $\theta_1 = (\beta,\gamma) = \varphi$. Later on we shall also consider $\varphi = f(\theta_1) = \beta$ or $\varphi = \gamma$. We successively analyze several BPTV's, *i.e.* conditional distributions of $(\beta,\gamma \mid b, M_0)$ and always retain the same prior distributions as in (3.3) and (3.4). Consequently, for each model, the corresponding posterior distributions are :

$$(b \mid X, M_0) \sim N(b_*, \sigma^2 h_*^{-1}) \tag{3.14}$$

$$\Big[\begin{pmatrix}\beta\\ \gamma\end{pmatrix} \mid X, M_1\Big] \sim N\Big[\begin{pmatrix}\beta_*\\ \gamma_*\end{pmatrix}, \sigma^2 H_*^{-1}\Big] \tag{3.15}$$

where, as usual (see *e.g.* Zellner (1971) :

$$h_* = h_p + z_1' z_1 \qquad b_* = h_*^{-1}(h_p b_p + z_1' y) \qquad H_* = H_p + Z'Z$$

$$\begin{pmatrix}\beta_*\\ \gamma_*\end{pmatrix} = H_*^{-1}\Big[H_p \begin{pmatrix}\beta_p\\ \gamma_p\end{pmatrix} + Z'y\Big] \ .$$

Under the coherent prior specification (3.5)-(3.6), it may be shown, either through clumsy analytical manipulations or through a simple argument on conditioning on the parameter space (see EBS section 1.4.3.) that one also has that : $(b \mid X, M_0) \sim (\beta \mid \gamma = 0, X, M_1)$, $b_* = E(\beta \mid \gamma = 0, X, M_1)$ and $\sigma^2 h_*^{-1} = V(\beta \mid \gamma = 0, X, M_1) = \sigma^2\ h_{11,*}^{-1}$.

For illustrative purposes, let us concentrate the attention on the class of normal BPTV's, *i.e.* :

$$\Big[\begin{pmatrix}\beta\\ \gamma\end{pmatrix} \mid b, M_0\Big] \ \sim \ N\big[r_0 + r_1 b, \sigma^2 W\big] \ . \tag{3.16}$$

where $r_0, r_1 \in I\!R^2$ and W is a 2×2 SPDS matrix. This specification implies :

$$\Big[\begin{pmatrix}\beta\\ \gamma\end{pmatrix} \mid X, M_0\Big] \ \sim \ N\big[r_0 + r_1 b_*, \sigma^2(h_*^{-1} r_1 r_1' + W)\big] \ . \tag{3.17}$$

For the discrepancy function $d(X)$ let us consider the information theoretic divergence *i.e.* $D(p,q) = \int ln\ \frac{p(x)}{q(x)} p(x)\ dx$. The discrepancy between $p^0(\beta, \gamma \mid X)$ and $p^1(\beta, \gamma \mid X)$ is accordingly evaluated for the case of non-singular variances as :

$$
\begin{aligned}
(3.18) \qquad d(X) &= D(p^0(\beta,\gamma \mid X), p^1(\beta,\gamma \mid X)) \\
&= \frac{1}{2}\big[ln \mid H_* \mid^{-1} - ln \mid W + h_*^{-1} r_1 r_1' \mid + tr H_* (W + h_*^{-1} r_1 r_1') \\
&+ \frac{1}{\sigma^2} \big(r_0 + r_1 b_* - \begin{pmatrix} \beta_* \\ \gamma_* \end{pmatrix}\big)' H_* \big(r_0 + r_1 b_* - \begin{pmatrix} \beta_* \\ \gamma_* \end{pmatrix}\big) - 2\big] \ .
\end{aligned}
$$

This statistic is to be calibrated against the predictive distribution under M_0. In this case, this is a normal distribution with parameters specified in (3.8) and $d(X)$ has the structure of a polynomial in y of degree 2 : $u_0 + u_1' y + y' U_2 y$; the predictive distribution of $d(X)$ is therefore the distribution of the sum of a constant, a linear combination of normal variates and a linear combination of χ^2 variates ; note that $u_1' y$ is not independent of $y' U_2 y$ but $u_1 = 0$ if r_0 is specified as :

$$
(3.19) \qquad r_0 = H_*^{-1} H_p (\beta_p,\ \gamma_p)' - r_1 h_*^{-1} h_p b_p,
$$

or case to be met later on.

Bayesian pseudo-true value degenerate at the classical one.

(i) Let us now consider some particular specifications of the BPTV in the class (3.16). A first case is a distribution degenerate at the classical pseudo-true value $(b\ 0)'$. This corresponds to :

$$
(3.20) \qquad r_0 = 0 \quad r_1 = (1 \qquad 0)' \quad W = 0
$$

in which case (3.17) becomes :

$$
(3.21) \qquad \big[\begin{pmatrix} \beta \\ \gamma \end{pmatrix} \mid X, M_0 \big] \ \sim \ N\big[\begin{pmatrix} b_* \\ 0 \end{pmatrix},\ \sigma^2 \begin{pmatrix} h_*^{-1} & 0 \\ 0 & 0 \end{pmatrix} \big]
$$

Because (3.21) is a degenerate distribution, the information-theoretic discrepancy between the joint posterior distributions of $(\beta,\ \gamma)'$ does not anymore provide a suitable statistic for diverging to infinity for any data X. One may use another discrepancy function, such as a Prohorov distance or a Kolmogorov- Smirnov distance.

(ii) If γ is the only parameter of interest - *i.e.* in the notation of section 2.4, $\varphi = f(\theta_1) = \gamma$ - the comparison between $(\gamma \mid X, M_0) \sim \delta_{\{0\}}$ (the Dirac measure on

0)and $(\gamma \mid X, M_1) \sim N(\gamma_*, \sigma^2\sigma^*_{22})$ (where $\Sigma^* = [\sigma^*_{ij}] = H_*^{-1}$) can be done either in terms of some distance among probability measures, as above, or, more naturally, in terms of a distance among expected values. A natural such distance would be :

$$d(X) = \frac{\gamma_*^2}{\sigma^2\sigma^*_{22}} \ . \tag{3.22}$$

Again this statistic is easily calibrated against $p^0(y \mid Z)$ for being, in this case, a general quadratic form in normal variates. When b and β are identified through the degenerate BPTV $\delta_{\{(b,0)'\}}$ and the prior specification satisfies $(b \mid M_0) \sim (\beta \mid \gamma = 0, M_1)$, one may also look at (3.22) as a function of two variables : the data X and the particular value of the parameter being tested (here $\gamma = 0$). Thus $d(X) = \rho(0, X)$ where :

$$\rho(\gamma, X) = \frac{(\gamma_* - \gamma)^2}{\sigma^2\sigma^*_{22}} \ . \tag{3.23}$$

Now $\rho(\gamma, X)$ is viewed as a measure of discrepancy between a particular hypothesis on γ and a particular data X. Florens and Mouchart (1989) have sketched a rather general approach based on that kind of function illustrated through this example. Identifying b and β means that $p^0(y \mid Z)$ may also be written as $p(y \mid Z, \gamma = 0)$ in the full Bayesian model characterized by a joint probability $p(\beta, \gamma, y \mid Z)$ and similarly $p^1(\gamma \mid X)$ becomes simply $p(\gamma \mid X)$. Now, $d(X) = \rho(0, X)$ may also be calibrated against that posterior distribution $p(\gamma \mid X)$. This is indeed in the spirit of Wald testing and also of evaluating a Bayesian p-value $\alpha_B(X)$ in the sense :

$$\alpha_B(X) = P^1[\rho(\gamma, X) \geq \rho(0, X) \mid X] \tag{3.24}$$

or a posterior confidence interval in the sense of (see *e.g.* Zellner (1971)) :

$$P[\mid \gamma - \gamma_* \mid \leq \mid \gamma - 0 \mid \mid X] \tag{3.25}$$

But $d(X) = \rho(0, X)$ may also be calibrated against the marginal joint distribution $p(\gamma, y \mid Z)$. The general idea of this approach is to consider both the particular value γ_0 for an hypothesis on γ and a particular value X_0 of data $X = (y, Z)$ as two pieces of information (conditionally on Z), to measure their conformity with their distribution $p(\gamma, y \mid X)$ through a discrepancy function $\rho(\gamma, X)$ and to calibrate $\rho(\gamma_0, X_0)$ against $p(y \mid Z_0, \gamma_0), p(\gamma \mid Z_0, y_0)$ or $p(\gamma, y \mid Z_0)$ according to whether the particular value γ_0, or y_0, or γ_0 and y_0 are deemed to be questionable.

(iii) Suppose now that β is the only parameter of interest. Under the normal prior specification in M_0 and in M_1-see (3.3) and (3.4)- and a normal BPTV -see (3.16)- the two posterior distributions of β are also normal, namely

$$(\beta \mid X, M_0) \sim (b \mid X, M_0) \sim N(b_*, \sigma^2 h_*^{-1}) \tag{3.26}$$

$$(\beta \mid X, M_1) \sim N(\beta_*, \sigma^2 h_*^{11}) \ , \tag{3.27}$$

where h_*^{11} is the $(1,1)$ element of H_*^{-1}. Thus the discrepancy between $p^0(\beta \mid X)$ and $p^1(\beta \mid X)$ may be evaluated exactly as in (3.18), namely

$$\begin{aligned} d(X) &= D(p^0(\beta \mid X), p^1(\beta \mid X)) \\ &= \frac{1}{2}[ln\ h_*^{11} - ln\ h_*^{-1} + h_*^{-1}(h_*^{11})^{-1} + (\sigma^2 h_*^{11})^{-1}(b_* - \beta_*)^2 - 1] \ . \end{aligned} \tag{3.28}$$

Note that under the coherent prior specification (3.5)-(3.6), we also have that :

$$(b_* - \beta_*)^2 = [h_*^{12}(h_*^{22})^{-1}\gamma_*]^2 \ . \tag{3.29}$$

Putting together (3.28) and (3.29) one obtains a Bayesian version of Hausman (1978) test, namely a test statistic based on a comparison of the inference on the parameter of interest β under the null ($p^0(\beta \mid X)$) and under the alternative hypothesis ($p^1(\beta \mid X)$) but the prior coherence implies that this is also a comparison of a marginal distribution on β ($p^1(\beta \mid X)$) and the conditional posterior distribution given the null hypothesis $p^0(\beta \mid X) = p^1(\beta \mid X,\ \gamma = 0)$. Analogously to the original Hausman test, $d(X) > 0$ even if $\gamma_* = E(\gamma \mid X, M_1) = 0$ and $d(X) = 0$ if and only if β and γ are a posteriori uncorrelated in Model 1, *i.e.* $h_*^{12} = 0$ and $h_*^{11} = h_*^{-1}$.

It should be noted that H_* and h_* involve both the prior specification and the value of the exogenous variable ; a similar property has been noticed in a sampling theory framework by . As in the previous case, $d(X)$ is distributed, in M_0, as a general quadratic form of normal variates. More details on the properties of that expression are given in Florens and Mouchart (1989).

Expected posterior distributions used as BPTV.

A non-degenerate sample-based BPTV may be constructed by taking the sampling expectation, under M_0, of the posterior distribution of θ_1 under M_1 :

$$p^0(\beta, \gamma \mid b, Z) = \int p^1(\beta, \gamma \mid X)\ p^0(y \mid Z, b)\ \ dy \tag{3.30}$$

Embedding the two distributions of the r.h.s. into a fictitious probability on $(\beta, \gamma, y \mid Z, b)$ with the condition of independence between (β, γ) and y given (Z, b), allows one to make use of standard manipulations on normal distributions, and to derive a normal distribution on $(\beta, \gamma \mid b, Z, M_0)$ with moments :

$$E(\beta, \gamma \mid b, Z, M_0) = H_*^{-1}[H_p \begin{pmatrix} \beta_p \\ \gamma_p \end{pmatrix} + Z' z_1 b] \tag{3.31}$$

$$V(\beta, \gamma \mid b, Z, M_0) = \sigma^2[H_*^{-1} + H_*^{-1} Z' Z H_*^{-1}] \ . \tag{3.32}$$

So one obtains a particular case of (3.16) with :

$$\begin{cases} r_0 & = H_*^{-1} \ H_p \begin{pmatrix} \beta_p \\ \gamma_p \end{pmatrix} \\ r_1 & = H_*^{-1} Z' z_1 \\ W & = H_*^{-1} + H_*^{-1} Z' Z H_*^{-1} = H_*^{-1}(H_* + Z'Z) H_*^{-1} \end{cases} \tag{3.33}$$

Note that for large n, and under rather general assumption on the behaviour of $Z_n' Z_n$ -such as $(Z_n' Z_n)^{-1} \to 0$ or $\frac{1}{n} Z_n' Z_n \to Q > 0$ – (r_0, r_1, W_n) may be approximated by $(0, (1,0)', 2(Z_n' Z_n)^{-1})$ and therefore $W_n \to 0$ as $n \to \infty$. Therefore this BPTV converges to a point mass on $(b, 0)'$. Thus, in finite sample, a discrepancy function $d(X)$ is evaluated as a particular case of (3.18) and is similarly calibrated against $p(y \mid Z, M_0)$. Note that a coherent prior specification does not affect (3.33) because (3.29) does not rely on the prior specification under M_0.

Optimal BPTV

Another strategy for specifying a BPTV is to look for a transition of probability that minimizes the predictive expectation, under M_0, of a specified discrepancy function $d(X)$. Finding the general solution to this problem is a difficult question

in the calculus of variation but restricted optimum are often easily obtained. We now give an example of such a situation.

When the discrepancy function is the information theoretic divergence, the optimum BPTV is easily obtained in the class (3.16) of gaussian BPTV when the prior specification is also normal as in (3.5)-(3.6). Indeed, in such a case, we have seen in (3.18) that $d(X)$ is a general quadratic form in y :

$$d(X) = u_0 + u_1' y + y' U_2 y \tag{3.34}$$

where u_0, u_1 and U_2 are functions of the parameters of the BPTV, *viz* r_0, r_1 and W, of the value of the exogenous variables Z and of the parameters of the prior distribution. As shown in Florens (1990), the solution of

$$\inf_{(r_0,r_1,W)} E^0[d(X)] \tag{3.35}$$

is given by :

$$\begin{cases} r_0^* & = H_*^{-1}\big[H_p \begin{pmatrix} \beta_p \\ \gamma_p \end{pmatrix} - h_p \; b_p \begin{pmatrix} 1 \\ c \end{pmatrix}\big] \\ r_1^* & = h_* H_*^{-1} \begin{pmatrix} 1 \\ c \end{pmatrix} \\ W^* & = H_*^{-1}\big[H_* - h_* \begin{pmatrix} 1 & c \\ c & c^2 \end{pmatrix}\big] H_*^{-1} = H_*^{-1} - h_*^{-1} r_1^* (r_1^*)' \end{cases} \tag{3.36}$$

where $c = (z_2' z_1)(z_1' z_1)^{-1}$ and provided that the term between brackets in W^* is SPDS, a condition to be commented upon later on. Note that the specification (3.36) satisfies condition (3.19) and therefore $u_1 = 0$ under an optimal BPTV. Asymptotically, under the usual conditions on $Z_n' Z_n$, we have :

$$(r_{0,n}^*, r_{1,n}^*, W_n^*) \to (0, (1,0)', 0) \tag{3.36}$$

and therefore the optimal BPTV also converges to a point mass on $(b, 0)'$. Under the optimal BPTV, the discrepancy function may be written as :

$$\begin{aligned} d^*(X) &= \frac{1}{2\sigma^2}(y' M_{z_1} Z H_*^{-1} Z' M_{z_1} y) \\ &= \frac{1}{2\sigma^2} h_*^{22} (y' M_{z_1} z_2)^2 \end{aligned} \tag{3.38}$$

where h_*^{22} is the $(2,2)^{\text{th}}$ element of H_*^{-1}. Note that under M_0, $\sigma^{-1} M_{z_1} y \sim N(0, I_{(n)})$ and therefore $d^*(X)$, under M_0, is distributed as a linear combination of independent $\chi^2_{(1)}$ variates. Note also that for large n, h_*^{22} may be approximated by $z_2' M_{z_1} z_2$ and $d^*(X)$ is approximatly equal to a Wald test statistics.

Let us have a closer look on the term between brackets in W_* appearing in (3.36) :

$$H_* - h_* \begin{pmatrix} 1 & c \\ c & c^2 \end{pmatrix} = H_p - h_p \begin{pmatrix} 1 & c \\ c & c^2 \end{pmatrix} + z_2' M_{z_1} z_2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.39}$$

Clearly this term is non negative only if the (1,1) element of H_p is not smaller than h_p, *i.e.* $V(\beta \mid \gamma, M_1) \leq V(b \mid M_0)$ and this condition is also sufficient for large n

because asymptotically $z_2' M_{z_1} z_2$ diverges to infinity whereas c tends to a limit. For a coherent prior specification, that condition is met through equality. Note also that when (3.39) is not SPDS, the optimum value of W is not given by (3.36) and should by obtained through a constrained optimization ; such would be the case when the prior specification is non-informative under M_1 (*i.e.* $H_p = 0$) of for some value of z_2 in finite sample.

3.3. Testing for unit roots

Recent econometric literature has paid considerable attention to a proper treatement of non-stationary time-series. At the level of modelling, most effort is concerned with the extraction of a stationary component, either a conditional one, though the introduction of non-stationary regressors (trend and/or exogenous variables) or, a marginal one, though differencing up to a suitable order (or, more generally, looking for cointegrated components). Nelson and Plosser (1982) has been path breaking in this field by calling the attention on the empirical relevance of the problem.

The discussion adopted a somewhat more controversial tone when it arrived at comparing Bayesian and sampling theory methods and became definitely hot when this comparison considered the problem of testing for unit root. A recent issue of the *Journal of Applied Econometrics* (vol.6, October-december 1991) offered a forum for exposing conflicting views, introduced by a "critical" paper by P.C.B Phillips (1991) along with his response to eight comments *viz.* Koop and Steel (1991), Leamer (1991), Kim and Maddala (1991), Poirier (1991), Schotman and Van Dijk (1991,b), Stock (1991), De Jong and Whiteman (1991) and Sims (1991). In a sense, that themed issue provides an interesting and up-dated "State of the Art" reminding past arguments and supplying new ones, in particular in terms of new empirical evidence on both simulated and actual data. Other recent contributions also include Schotman and Van Dijk(1991,a) and hitherto unpublished papers by Lubrano.

As often happened in such cases, the temperature of the discussion is raised by mixing several ingredients. Some of those ingredients are of general nature such as questions about the foundations and relevance of Bayesian and of sampling theory methods, the relevance of the problems raised by Bayesian inference under improper prior specification, the numerical problems such as the relative merits of the Laplace approximation of posterior distributions, the possibly striking difference between sampling characteristics, such as p-value, and the posterior characteristics, such as posterior odds or Bayes factors and, last but not least, the use of affectionate qualifications such as "sterile ideas", "unreasonable", "logically unsound" or "wrong headed and

unlightening". One ingredient, at least, is rather particular to the topic, and possibly explains the level of the temperature, namely that around the unit roots the likelihood is smooth in finite sample but asymptotically discontinuous: this issue has been raised in Dickey and Fuller (1979), see also Phillips (1987, 1991 a, 1991 b). Thus Bayesian methods, being essentially of a small sample nature, may be expected to diverge from asymptotic sampling methods ; also Bayesian inference based on improper prior specification, justifiable (when possible !) as providing asymptotic approximation, may also be expected to raise the controverse generally associated to the use of such prior specifications.

Clearly we can't do justice of all arguments in the framework of this survey. The main ideas may nevertheless be sketched by discussing the simplest case, namely the univariate AR(1) case. Denoting $y_i^j = (y_i,\ y_{i+1}, \cdots, y_j)$ for any $i < j$, let us consider :

$$(y_t \mid y_0^{t-1}) \sim N(\rho y_{t-1}, \sigma^2) \qquad 1 \leq t \leq T \ . \tag{3.40}$$

In this dynamic context, observations are naturally indexed by t and T becomes the sample size. In what follows, we only consider models conditional to y_0, which is therefore not modelled, without discussing under which circunstances is such a conditioning admissible. Thus the likelihood function is derived form :

$$p(y_1^T \mid y_0, \rho, \sigma^2) = (2\pi)^{-(T/2)} \quad \sigma^{-T} \exp -\frac{1}{2\sigma^2} \sum_{1 \leq t \leq T} (y_t - \rho y_{t-1})^2 \ . \tag{3.41}$$

From a sampling point of view, the asymptotic distribution associated to the sequence $p(\hat{\rho}_T \mid y_0, \rho, \sigma^2)$ - where $\hat{\rho}_T$ is the MLE of ρ based on y_0^T - is terribly different when $\rho < 1$, $\rho = 1$ or $\rho > 1$. This feature possibly produces disconnected (and therefore non convex) confidence regions. From a Bayesian point of view, a particular value of ρ is irrelevant as far as, once a prior distribution is assigned to (ρ, σ), the evaluation of the posterior distribution will run smoothly for any finite sample size, possibly calling for simple numerical integrations. However, for testing the hypothesis $\rho = 1$, the pathologies of the likelihood function implies a particular sensitivity to the prior specification, (See Sims and Uhlig(1991)). We shall first discuss the prior specification and next the choice of a particular testing procedure.

Considering the decomposition (2.17)-(2.18) of the prior distribution, we first discuss the specification of $p(\rho, \sigma \mid \rho \neq 1)$. In general, one finds in the literature prior

distributions of the form :

$$p(\rho, \sigma \mid \rho \neq 1) \propto \frac{1}{\sigma} g_1(\rho) \tag{3.42}$$

with no discussion of the specification relative to σ. The alternative hypothesis is implicitly defined by the support of the second component. One finds at least three such specifications : (i) Sims (1988) employs $g_1(\rho) = \mathbb{1}_{\mathbb{R}}(\rho)$, *i.e.* the density of Lebesgue measure on the real line, therefore excluding no value of ρ and accepting eventually explosive models. Remember that Lebesgue measure favours extreme values in the sense that the measure of the interval $(-1, +1)$ relative to the measure of $(-a, +a)$ tends to zero as a tends to infinity (ii) Schotman and Van Dijk (1991,a) starts with a uniform proper distribution of ρ on the bounded interval $[a, 1]$ *i.e.* $g_1(\rho) = \frac{1}{1-a} \mathbb{1}_{[a,1]}(\rho)$ (with $-1 < a < 1$) but leaves the component of σ improper. Thus their alternative hypothesis is $[a, 1]$ *i.e.* a stationary model (a priori *a.e.*). They show that the result of testing ρ crucially depends on the value of a and suggest an empirical procedure for specifying a and develop further their analysis in (1991b). (iii) Phillips (1991a) criticizes the case of flat prior for favouring unduly the stationary use and argues instead for the case of Jeffrey's non informative prior *i.e.* $g_1(\rho) = (1-\rho^2)^{-1/2}\{T-(1-\rho^{2T})[(1-\rho^2)^{-1}-\sigma^{-2}y_0^2]\}^{1/2} \mathbb{1}_{\mathbb{R}}(\rho)$, taking again the real line as alternative hypothesis. He also uses the Laplace approximation for analyzing the posterior distribution and reinterprets the series analyzed by Nelson and Plosser (1982).

As far as the choice of a testing procedure is concerned, two approaches have been proposed. A first one, in line with Lindley (1961), starts with a smooth prior specification, *i.e.* $P(\rho = 1) = 0$ and evaluates the posterior probability either of regions around 1 with the idea of asking whether the unit root lies in a high enough posterior probability region as in Phillips (1991a) or of region of non stationarity such as $\{\rho > 1\}$ or $\{\rho < 1\}$ as in Phillips (1991a), Kim and Maddala (1991) or Schotman and Van Dijk (1991b). The advantage of that class of procedures is to make the economy of eliciting a prior probability of the null hypothesis and to be well defined even under improper prior specification. A second approach, maybe more traditional in the Bayesian establishment, evaluates Bayes factor, posterior odds or posterior probability of hypotheses. As mentionned in section 2.2.1. this class of procedures requires a proper prior specification along with a strictly positive mass on the null hypothesis ; Schotman and Van Dijk (1991,a and b) report problems issued from overlooking this requirement.

Extending model (3.40) to more complex structures raises new difficulties. For instance when $y_t = \mu + u_t$ with $u_t = \rho u_{t-1} + \varepsilon_t$ and $\varepsilon_t \sim I.N(0, \sigma^2)$, a local identificaton problem appears at the null hypothesis $\rho = 1$. This issue, similar to the rank condition in simultaneous equation model, see Madalla (1976) and Dräze and Mouchart (1990), is analyzed in the present context by Schotman and Van Dijk (1991 b) under a partially non informative prior specification. The same paper, along with Phillips (1991 a), also considers an intercept term and a linear trend in order to analyze further on problems of transient dynamics or of testing trend stationarity against difference stationarity.

4. Other contributions

One of the first issues to be faced by the author(s) of a survey is to border the field to be surveyed. In this survey we put much emphasis on two ideas. Firstly hypothesis testing and model choice have been dealt with as a single class of problems met with so strikingly varied motivations that no clear distinction among them seems to be operationally fruitful. Secondly Bayesian thinking is rich enough to accomodate to that variety of situations and is much more flexible that a mechanic prior-to-posterior transformation ; in particular, the predictive distributions have been shown to play an important role for this class of problems.

Putting emphasis on general inference, on model label or on parameters of interest, has lead to put less emphasis on solving specific decision problems. Thus strictly decision oriented procedures have at time been alluded to but not dealt with in any systematic way. This is in particular, the case of model choice procedures based on specific criteria such as AIC or BIC with different degrees of involvment with Bayesian idea. This literature is rather vast but somewhat disconnected from the main theme considered in this survey. The interested reader may consult Schwartz (1978) or Zellner (1978).

Section 3 has treated two exemplary fields of application but many more have been considered in the literature. A notable one is the testing of exogeneity considered by Florens and Mouchart (1989), Lubrano, Pierse and Richard (1986), Steel and Richard (1991). Many contributions have also dealt with the problem of specification testing in linear models, as in Bauwens and Lubrano (1991). A long list of applied works using Bayesian testing has been omitted. We just quote as an example a recent issue of the *Journal of Econometrics* (see *e.g.* Connolly (1991), Koop (1991), Mc Culloch and Perossi (1991) or Moulton (1991)).

REFERENCES

BAUWENS, L. and M. LUBRANO (1991), "Bayesian Diagnostics for Heterogeneity", *Annales d'Economie et de Statistique*, **20-21**, 17-40.

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

BERGER, J.O. (1986), "Are P-Values Reasonable Measures of Accuracy", in : *Pacific Statistical Congress*, I.S. Francis et al., eds., Amsterdam : North Holland.

BERGER, J.O. and D. BERRY, (1988), "Statistical Analysis and the Illusion of Objectivity", *The American Scientist*, **76**, 159-165.

BERGER, J.O. and M. DELAMPADY (1987), "Testing Precise Hypotheses", *Statist. Sci.*, **2**, 317-352.

BERGER, J.O. and T. SELLKE (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p-Values and Evidence", *J. Amer. Statist. Ass.*, **82**, 112-122.

BERK, R.H. (1966), "Limiting Behavior of Posterior Distributions when the Model is Incorrect", *The Annals of Mathematical Statistics*, **37**, 51-58.

BERNARDO, J.M (1980), "A Bayesian Analysis of Classical Hypothesis Testing" in : *Bayesian Statistics* (J.M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith, eds.), Valencia : The University Press, 605-618.

CASELLA, G. and R.L. BERGER (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem", *J. Amer. Statist. Ass.*, **82**, 106-111.

CONNOLLY, R.A., (1991), "Posterior Odds Analysis of the Weekend Effect", *J. Econometrics*, **49**, 51-104.

DEGROOT, M.H. (1970), *Optimal Statistical Decisions.* New York: McGraw-Hill.

DEGROOT, M.H. (1973), "Doing what Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio", *J. Amer. Statist. Ass.*, **68**, 966-969

DE JONG, D.N. and C.H WHITEMAN (1991), "The Case for Trend- Stationarity in Stronger than we Thought", *J. Appl. Econometrics", **6**, 413-422.*

DICKEY, D.A. and W.A. FULLER (1979), "Distribution of the Estimators for Autoregressive Time Series with a Unit Root" *J. Amer. Statist. Ass.*, **74**, 427-451.

DICKEY, J.M. (1975), "Bayesian Alternatives to the F-test and Least-Squares Estimates in the Normal Linear Model", in: *Studies in Bayesian Econometrics and Statistics in Honour of L.J. Savage* (S.E. Fienberg and A. Zellner, eds.), Amsterdam: North-Holland.

DICKEY, J.M. (1977), "Is the Tail Area Useful as an Approximate Bayes Factor ?" *J. Amer. Statis. Ass.*, **72**, 138-142.

DICKEY, J.M. and J. KADANE (1980), "Bayesian Decision Theory and the Simplification of Models" in *Evaluation of Econometric Models*, (J. Kmenta and J. Ramsey, eds.), Academic Press, 245-268

DREZE, J. and M. MOUCHART, (1990), "Tales of Testing Bayesians" in *Contribution to Economic Theory* (R.A.L. Carter, J. Dutta and A. Ullah, eds.), New-York : Springer Verlag, 345-366.

FLORENS, J.P. (1990), "Parameter Sufficiency and Encompassing", in : *Essays in Honor of Edmond Malinvaud*, Cambridge : The MIT Press (1990), **3**, 115-136.

FLORENS, J.P. and M. MOUCHART (1985), "Model Selection : Some Remarks from a Bayesian Viewpoint", in :*Model Choice*, (J.P. Florens, M. Mouchart, J.P. Raoult and L. Simar, eds.), Bruxelles : Publications des FacultÇs Universitaires Saint-Louis, 27-44.

FLORENS, J.P. and M. MOUCHART (1989), "Bayesian Specification Test", in : *Contributions to Operations Research and Economics*, (B. Cornet and H. Tulkens, eds), Cambridge : The MIT Press, 467-490.

FLORENS, J.P., M. MOUCHART and J.M. ROLIN (1990), *Elements of Bayesian Statistics*, New-York : Marcel Dekker.

FLORENS, J.P., M. MOUCHART and S.SCOTTO (1983), "Approximate Sufficiency on the Parameter Space and Model Selection", in : *44th Session of the International Statistical Institute : Contributed Papers*, **2**, 763-766.

GAVER, K.M. and M.S. GEISEL (1974), "Discriminating Among Alternative Models: Bayesian and Non-Bayesian Methods". in: *Frontiers in Econometrics.*, (P. Zarembka, ed.), New York: Academic Press, 49-77.

GEISSER, S. and W.F. EDDY (1979), "A Predictive Sample Approach to Model Selection", *J. Amer. Statis. Ass.*, **74**, 153-160.

GOOD, I.J. (1985), "Weight of Evidence : A Brief Survey ", in : *Bayesian Statistics 2*, (J.M. Bernado, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.), Amsterdam : North Holland.

HAUSMAN, J.A. (1978), "Specification Tests in Econometrics", *Econometrica*, **46**, 1251-1272.

HILL, B. (1982), "Comment on Lindley's Paradox, by G. Shafer", *J. Amer. Statis. Ass.*, **77**, 344-347.

HILL, B. (1985-1986), "Some Subjective Bayesian Considerations in the Selection of Models", *Econometric Reviews*, **4**, 191-282

HILL, B. (1990), "A Theory of Bayesian Data Analysis", in *Bayesian and Likelihood Methods in Statistics and Econometrics*, S. Geisser, J.S. Hodges, S.J. Press and A. Zellner (eds.)), Amsterdam: Elsevier Science Publishers B.V. (North Holland), 49-73.

HODGES, J.S. (1990), "Can/May Bayesian Do Pure Tests of Significance ?", in *Bayesian and Likelihood Methods in Statistics and Econometrics*, S. Geisser, J.S Hodges, S.J. Press and A. Zellner (eds.)), Amsterdam: Elsevier Science Publishers B.V. (North Holland), 75-90.

HWANG, J.T. C. Robert, M.T. Wells and R.H. Farrel (1989), "Estimating Acceptance Probabilities", Mimeo, Mathematical Sciences Institute, Cornel University, Itaka, (USA), to appear in Ann. Statist..

JEFFREYS, H. (1967), *Theory of Probability*, 3rd rev. ed., London : Oxford University Press.

KIM, I.M. and G.S. MADDALA (1991), "Flats Priors vs. Ignorance Priors in the Analysis of the AR(1) Model", *J. Appl. Econometrics*, **6**, 375-380.

KLEIN, R.W. and S.J. BROWN (1984), "Model Selection when there is Minimal Prior Information", *Econometrica*, **52**, 1291-1312.

KOOP, G. (1991), "Cointegration Tests in Present Value Relationships : A Bayesian Look at the Bivariate Properties of Stock Prices and Dividends", *J. Econometrics*, **49**, 105-140.

KOOP, G. and M.F.J. STEEL (1991), A Comment on "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends" by P.C.B. Phillips, *J. Appl. Econometrics*, **6**, 365-370.

LEAMER, E.E. (1978), *Specification Searches: Ad Hoc Inferences with Non-experimental Data*, New York: John Wiley.

LEAMER, E.E. (1983,b), "Model Choice and Specification Analysis" in: *Handbook of Econometrics*, (Z. Griliches and M.D. Intriligator eds.), **1**, Amsterdam : North-Holland.

LEAMER, E.E. (1991), Comment on "To Criticize the Critics", *J. Appl. Econometrics*, **6**, 371-374.

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

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

LINDLEY, D.V. (1961), "The Use of Prior Probability Distributions in Statistical Inference and Decisions", *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probabability*, **1**, Berkeley : University of California Press, 453-468.

LINDLEY, D.V. (1971), *Bayesian Statistics, A Review*, Regional Conference Series in Applied Mathematics, S.I.A.M.

LINDLEY, D.V. (1977), "A Problem in Forensic Science", *Biometrika*, **64**, 207-213.

LUBRANO, M., R.G. PIERSE and J.F. RICHARD (1986), "Stability of a U.K. Money Demand : A Bayesian Approach to Testing Exogeneity", *R. Economic Studies*, **53**, 603-634.

MADDALA, G.S (1976), "Weak Priors and Sharp Posteriors in Simultaneous Equation Models", *Econometrica*, **44**, 345-351.

MC. CULLOCH, R and PEROSSI (1991), "A Bayesian Approach to Testing the Arbitrage Pricing Theory", *J. Econometrics*, **49**, 141-168.

MORENO. E. and J.A. CANO (1989), "Testing a Point Null Hypothesis: Asymptotic Robust Bayesian Analysis with Respect to Priors Given on a Subsigma Field", *Int. Statist. Rev.*, **57**, 221-232.

MOULTON, B.R. (1991), "A Bayesian Approach to Regression Selection and Estimation, with Application to a Price Index for Radio Services", *J. Econometrics*, **49**, 169-194.

NELSON, C.R. and C. PLOSSER (1982), "Trends and Random Walks in Macroeconomic Time Series : Some Evidence and Implications", *Journal of Monetary Economics*, **10**, 138-162.

PHILLIPS, P.C.B. (1987), "Time Series Regression with a Unit Root", *Econometrica*, **55**, 277-301.

PHILLIPS, P.C.B. (1991 a), "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends", *J. Appl. Econometrics*, **6**, 333-364.

PHILLIPS, P.C.B. (1991b), "Bayesian Routes and Unit Roots: De Rebus Prioribus Semper Est Disputandum", *J. Appl. Econometrics*, **6**, 435-474.

PICCINATO, L. (1990), "Sull'interpretazione del Livello di Significativita'Osservato", in *Scritti in Omaggio a Luciano Daboni*, Trieste: Edizioni Lint, 199-213.

POIRIER, D.J., (1985),"Bayesian Hypothesis Testing in Linear Models with Continuously Induced Conjugate Priors Across Hypotheses", in *Bayesian Statistics 2*,(J.M. Bernado, M.H. De Groot, D.V. Lindley and A.F.M. Smith, eds.) Amsterdam : North Holland, 711-722.

POIRIER, D.J., (1988), "Frequentist and Subjectivist perspectives on the Problems of Model Building in Economics", *Journal of Economic Perspectives*, **2**, 121-170.

POIRIER, D.J., (1991), "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends", *J. Appl. Econometrics*, **6**, 381-386.

PRATT, J., H. RAIFFA and R. SCHLAIFER (1964), "Foundations of Decisions under Uncertainty : an Elementary Exposition", *J. Amer. Statist. Ass.* **59**, 353-375.

RAFTERY, A.E. (1986), "A Note on Bayes Factors for Log-Linear Contingency Table Models with Vague prior Information", *J. R. Statist. Soc.*, **48**, 249-250.

RUBIN, H. (1971), "A Decision-Theoretic Aproach to the Problem ot Testing a Null Hypothesis", in *Statistical Decision Theory and Related Topics* (S.S. Gupta and J. Yackel, eds.)New York : Academic Press.

SAVAGE, L.J. (1954), *The Foundations of Statistics*. New-York : John Wiley.

SCHOTMAN, P. and H. VAN DIJK (1991,a), "A Bayesian Analysis of the Unit Root in Real Exchange Rates", *J. of Econometics*, **49**, 195-238.

SCHOTMAN, P. and H. VAN DIJK (1991,b), "On Bayesian Routes to Unit Roots", *J. Appl. Econometrics*, **6**, 387-402.

SCHWARZ, G. (1978), "Estimating the Dimension of a Model", *Ann. Statist.*, **6**, 461-464.

SHAFER, G. (1982), "Lindley's Paradox" (With discussion), *J. Amer. Statis. Ass.*, **77**, 325-351.

SHANKEN, Hay (1987), "A Bayesian Approach to Testing portfolio Efficiency, *Journal of Financial Economics*, **18**, 195-215.

SIMS, C.A. (1988), "Bayesian Skepticism on Unit Root Econometrics", *Journal of Economic Dynamics and Control*, **12**, 463-474.

SIMS, C.A. and H. UHLIG (1991), "Understanding Unit Rooters : A Helicopter Tour, *Econometrica*, 59, 1591-1600.

SIMS, C.A. (1991), Comment on "To Criticize the Critics" by P.C.B. Phillips, *J. Appl. Econometrics*, **6**, 423-434.

SPIEGELHALTER, D.J and A.F.m. SMITH (1982), "Bayes Factors for Linear and Log-linear Models with Vague Prior Information", *J. R. Statist. Soc.*, B, **44**, 377-387.

STEEL, M. and J.F. RICHARD (1991), "Bayesian Multivariate Exogeneity Analysis : An Application to UK Demand Equation", *J. Econometrics*, **49**, 239-274.

STOCK, J.H. (1991), "Bayesian Approaches to the 'Unit Root' Problem: A Comment", *J. Appl. Econometrics*, **6**, 403-412.

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

ZELLNER, A. (1978), "Jeffreys Bayes Posterior Odds Ratio and the Akaike Information Criterion for Discriminating between Models", *Economic Letters*, **1**, 337-342.

ZELLNER, A. (1984), "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results", in : *Basic Issue in Econometrics*, Chicago: University of Chicago Press, 275-304.

ZELLNER, A. and A. SIOW (1980), "Posterior Odds Ratios for Selected Regression Hypotheses", in: *Bayesian Statistics*, (J.M. Bernado, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds.) Valencia : University Press, 585-603. Reply, 638-643.