Bayesian Encompassing Specification Tests of a Parametric Model against a Nonparametric Alternative

J.P. Florens[1] J.F. Richard[2] J.M. Rolin[3]

June 15, 2000

[1] GREMAQ and IDEI, Universit¶e des Sciences Sociales de Toulouse, France.
[2] Department of Economics, University of Pittsburg, USA.
[3] Institut de Statistique and CORE, Universit¶e Catholique de Louvain, Belgium.

Abstract

Encompassing tests of a model M_0 are based upon the notion that M_0 ought to be able to account for results derived upon alternative models. Within a Bayesian framework, posterior distributions obtained under an alternative model M_1 are explicitely compared with posterior distributions obtained within M_0 by means of a suitable "transition" distribution for the parameters of M_1 conditionnally on those of M_0.

In our paper, we propose a generic Bayesian procedure for testing the capabilities of a parametric model M_0 to encompass a wide range of inferential results derived under a nonparametric alternative M_1. By combining the use of Dirichlet prior measures on M_1 with that of Monte Carlo simulation techniques, we develop exact operational and highly °exible Bayesian encompassing test procedures of M_0 relative to M_1. An application to an exponential lifetime model M_0 illustrates the performance of our procedure.

Aknowledgement

We are grateful to the editor and two four referees for exceptionally insightful comments that led to a greatly improved version of our initial paper. All remaining errors are ours.

1 Introduction

The objective of our paper is that of proposing a generic Bayesian procedure for testing the validity of a parametric model M_0. Specifically, we rely upon Monte Carlo simulation techniques to construct an exact Bayesian test of whether or not M_0 encompasses a general nonparametric model M_1.

As discussed by Florens et al (1996) Bayesian encompassing corresponds to a concept of sufficiency among models, dual to the traditional concept of sufficiency (among statistics) as defined by Blackwell (1951,1953) and Lecam (1964) and obtained from the latter by interchanging parameters and statistics. The model M_0 encompasses (Bayesian) inference results derived on M_1 if the latter can be reproduced within M_0 without any additional data treatment beyond that already necessitated by the analyis of M_0 itself. In other words, M_1 then becomes "inferentially redundant" relatice to M_0. In a wide range of scientific disciplines the capability of one's current model (or theory) to encompasses findings derived under other models is perceived to be a critical component of its validation.

Conversely, failure to encompass results obtained on M_1 constitutes prima facie evidence of "deficiencies" of M_0 which, if deemed relevant to one's objectives, would indicate a need for further refinements of M_0. It does not, however, warrant the conclusion that M_1 is to be preferred to M_0 since the requirement that M_0 ought to encompass rival models is by no means conditional upon the validity of the latter. Significant advances in sciences have resulted from the need to account for "anomalous" findings derived under invalid models and/or theories.

This latter comment illustrates a fundamental difference between an encompassing test, whereby the model M_1 is essentially instrumental in the construction of the test but generally does not constitute an acceptable alternative to M_0 itself, and a Bayesian model selection procedure whereby one chooses between two alternative models M_0 and M_1 which are generally treated symmetrically relative to one another (notwithstanding the obvious fact that loss function and prior probability assessments can be used to favor one model relative to the other). Actually, as we shall discuss further later, encompassing tests are based upon a comparison between posterior densities while model selection depends upon a ratio of predictive densities (Bayes factor, see e.g. Florens and Mouchart (1993)) This difference will turn out to be highly significant in the context discussed hereafter, whereby M_0 is a parametric model and M_1 a nonparameteric one - a situation which is aimed at providing a flexible and yet quite stringent test of the validity of M_0 by examining its capability to encompass a potentially broad range of results derived under a general nonparametric specification for M_1. In particular, the use of a natural conjugate prior Dirichlet process for M_1 - which assigns probability one to discrete measure but is, nevertheless, dense within a set of continuous measures - has critical and somewhat "extreme" implications for model selection while it produces a fully operational fundamentally well-behaved and highly flexible encompassing test procedure.

In particular, flexibility originates from the fact that our encompassing test procedure allows its user to select specific functionals of interest (moment, tail probabilities,...) whose posterior densities on M_1 have to be encompassed by M_0, i.e. to decide which findings associated with M_1 are particularly relevant to the proprietor of M_0 in any given (decisional) application. In constrast, Bayesian model selection procedures are based upon a single Bayes factor and often provide limited insights as to which specific aspects of the competing models drove the final decision.

Finally, we wish to mention here that the encompassing tests we propose prolong a long tradition in sampling statistics, initiated at the turn of the century by K. Pearson with chi-squared test statistics, extended in the thirties with the Kolmogorov-Smirnov and Cramer von Mises test statistics and broadly labeled today as "goodness of fit" tests. Recent extensions are discussed e.g. by Revesz (1984). Additional related contributions are specifically discussed in the sequel of our paper after the relevant concepts and notation have been introduced.

Our paper is organized as follows : The models and notation are introduced in section 2; Posterior odds are derived in section 3; Encompassing is discussed in section 4 (principle in section 4.1. and nonparametric version in section 4.2, respectively); an application to the exponential distribution is presented in section 5 and section 6 concludes.

2 Models and notation

The observation $x = (x_1, \ldots, x_n)$ is an i.i.d. sample of a random variable X with support S in $\mathbb{R}^p$. Let M_0 denote the parametric model to be (in)validated and M_1 a nonparametric model to be encompassed. On M_0 all probability are assumed to be absolutely continuous with respect to an appropriate σ-finite dominating measure and are represented by the corresponding density functions. On M_1 we are dealing with probability measures on a sampling distribution F. In order to unify the presentation of our results and at the cost of an abuse of notation, we shall equally represent such probabilities by a "density function" on F. Densities on a (finite-dimensional) parameter θ, on a sample x and on a distribution F are denoted μ_0, p and ν, respectively. In addition, p and ν are subscripted with the index of the relevant model. Conditioning is represented in the usual way. The symbol » reads as "is distributed as".

The sampling distribution of X on M_0 is denoted by F^{μ} with parameter μ in £ ½ IR^k and with density $f(xj\mu)$. The sampling density of the sample x is then given by

$$p_0(xj\mu) = \mathop{\mathbf{Y}}_{i=1} f(x_ij\mu): \qquad (1)$$

Let ${}^1{}_0(\mu); p_0(x)$ and ${}^1{}_0(\mu jx)$ denote the prior density of μ, the predictive density of x and the posterior density of μ, respectively. Following Bayes theorem, we have

$${}^1{}_0(\mu):p_0(xj\mu) = {}^1{}_0(\mu jx):p_0(x): \qquad (2)$$

The sampling distribution of X on M_1 is denoted by F. Following Ferguson (1973), a natural conjugate prior distribution for F is provided by a Dirichlet measure which is parametrized by a real positive number n_a and a probability F_a on IR^p, say

$$F \text{ » } Di(n_aF_a): \qquad (3)$$

In view of its importance for the sequel of the discussion, we briefly discuss here the interpretation of the Dirichlet measure (3). Consider first a fixed partition $B = (B_1; :::; B_R)$ of the support S of F and let $\text{¼}^0 = (\text{¼}_1; :::; \text{¼}_R)$, with $\text{¼}_r = F(B_r)$, denote the corresponding vector of sampling probabilities. A sufficient statistic for the sample x is given by the vector $f^0 = (f_1; :::f_R)$, where f_r denotes the proportion of observations in B_r. The random variable nf has multinomial distribution with parameter(s) ¼ (and n). Its density function is given by

$$g(nfj\text{¼}) = n! \mathop{\mathbf{Y}}_{r=1} \frac{(\text{¼}_r)^{nf_r}}{(nf_r)!}: \qquad (4)$$

A natural conjugate distribution for ¼ is given by a Dirichlet distribution whose support is the simplex

$$S_R = f\text{¼}; \text{¼}_r > 0; \mathop{\mathbf{X}}_{r=1} \text{¼}_r = 1g \qquad (5)$$

of dimension $R ¡ 1$, with parameters $n_a > 0$ and $f_a \; 2 \; S_R$ and whose density function is given by

$$º_1(¼) = ¡(n_a) \prod_{r=1} \frac{(¼_r)^{n_a f_{ar} ¡ 1}}{¡(n_a f_{a;r})}: \tag{6}$$

The M_1-posterior distribution of ¼ is itself a Dirichlet distribution with parameters

$$n_p = n_a + n; \quad n_p f_p = n_a f_a + n f: \tag{7}$$

The Dirichlet prior measure in (3) is then defined as one which assigns a Dirichlet distribution to the random vector $F(B) = (F(B_1); ::::; F(B_R))$ associated with any arbitrary partition B of the support S of ¼. That is to say that the Dirichlet prior measure in (3) assigns the following density to $F(B)$:

$$º_1(F(B)) = ¡(n_a) \prod_{r=1} \frac{[F(B_r)]^{n_a F_a(B_r) ¡ 1}}{¡(n_a F_a(B_r))} \tag{8}$$

with support S_R. It follows from our discussion that the posterior distribution of F on M_1, is a Dirichlet measure with parameters

$$n_p = n_a + n; \quad n_p F_p = n_a F_a + n F_n; \tag{9}$$

where F_n is the empirical distribution of the sample

$$F_n = \frac{1}{n} \sum_{i=1} ±_{x_i}; \tag{10}$$

and $±_{x_i}$ denotes the Dirac measure at x_i which is such that $±_{x_i}(A) = 1\!\mathrm{I}_A(x_i)$ for A ½ $\mathrm{I\!R}^p$. Following Ferguson (1973), the predictive distribution $p_1(x)$ on M_1 can be represented in a variety of ways. The following sequential representation proves to be the most operational for our analysis

$$\begin{array}{c} x_1 \; » \; F_a \\ x_{i+1} j x_1; ::::; x_i \; » \; \frac{n_a}{n_a+i} F_a + \frac{i}{n_a+i} F_i; \text{ for } i : 1 \; ! \; n \; ¡ \; 1: \end{array} \tag{11}$$

Within an encompassing framework it may be desirable to achieve some degree of "coherency" between the prior distributions of the two models. The concept of coherency introduced by Florens et al (1996) is too strong in that it requires that the two priors be linked together by a transition which actually depends upon sample size. An alternative approach consists of choosing the prior parameter F_a in such a way that the two models be indistiguishable from one another on the basis of a single observation x_1 (i.e. that the Bayes factor associated with a single observation be equal to one). In view of the representation of the M_1 predictive in (11), this is equivalent to setting F_a equal to the former that is to say

$$F_a(B) = \int_B p_0(x_1)dx_1; \quad \forall B \in \mathcal{B}_p \tag{12}$$

Definition 2.1 The prior measure $Di(n_a F_a)$ is predictive-coherent if and only if condition (12) holds.

$\square$

Note that, independently of whether or not condition (12) holds, $F_a(B)$ represents the prior expectation of $F(B)$ on M_1. As usual within a natural conjugate framework and as illustrated by (9), n_a represents the prior "weight" attached to F_a in a metric which is directly comparable to the actual sample size n and is, therefore, often refered to as to a "hypothetical" sample size - see e.g. Raiffa and Schlaifer (1961).

Our motivation for selecting a prior Dirichlet measure on M_1 is fourfold : (i) It is easy to specify since, as we just discussed, its parameters have "natural" interpretations ; (ii) It is flexible in that, following Ferguson (1973) it generates a dense subset of probabilities ; (iii) It is natural conjugate so that its two parameters are revised by means of the simple (additive) convolution rule given by (9) ; and (iv), it will prove especially operational for the construction of simulation based encompassing test procedures.

Nevertheless, two important characteristics of the Dirichlet measure deserve additional comments. Firstly, Dirichlet measures generate discrete probabilities. Actually, this property is intrinsic to inference in nonparametric models since the empirical distribution which is itself discrete constitutes a sufficient statistic and, therefore, serves as the basis of numerous goodness of fit tests, such as the Kolmogorov-Smirnov test. One might consider instead using prior measures whose realizations are continuous. Several approaches have been explored in the recent literature. One consists of introducing prior measures whose trajectories are the convolution of a Dirichlet realization and a of preassigned continuous probability -see e.g. Lo (1984) or

Escobar and West (1995). From our viewpoint, the main drawback of that approach lies in the complexity of the resulting posterior distributions (for example, posterior expectations are linear combinations comprising as many terms as there are partitions of the set $\{1, \ldots, n\}$). An alternative approach consists of selecting a logistic normal measure -see e.g. Leonard (1978) or Lenk (1988, 1991). Here again posterior distributions are complicated and, furthermore, this prior specification produces Bayesian procedures which are not convergent. See also Lavine (1992, 1994) for an alternative way of generating continuous probabilities. Preliminary investigations suggest that there are no conceptual difficulties generalizing the encompassing test procedures which are developed here to such continuous extensions and, furthermore, that actual computation are likely to prove simpler than those necessitated e.g. to produce posterior odds.

Secondly, as illustrated by (11), predictive probabilities associated with Dirichlet prior measures are intrinsically "bumpy" in the sense that they generate ties with positive probabilities. As we shall discuss further below, this property has "extreme" implications for (Bayesian) model selection procedures which essentially reduce to a test of continuity versus discreteness. It also produces infinite divergence measures in a "brute force" comparison between posterior densities. On the other hand functionals of F admit posterior densities on M_1 under mild technical conditions and will, therefore, result in encompassing test procedures which are no longer bumpy.

3 Posterior odds

Let $\delta \in \{0, 1\}$ denote a binary model indicator. Let P denote a probability distribution on $(\delta; X)$ which is such that

$$P(\delta = 0) = a; \quad P(\cdot \mid \delta = i) = P_i; \tag{13}$$

where P_i denotes the predictive probability on M_i, for $i = 1; 2$. It is always possible to dominate P_0 and P_1 by a σ-finite measure such as, for example, $\frac{1}{2}(P_0 + P_1)$. Let, therefore, dP_i denote the density of P_i w.r.t. a suitable dominating measure[1]. The corresponding Bayes factor is then given by the ratio dP_0/dP_1 (not to be confused with the Radom-Nikodyn derivative of P_0 with respect to P_1 which does not exist in the present context). The posterior odds ratio is given by the product of the prior odds ratio $a/(1 - a)$ by the Bayes factor and the posterior probability of model M_0 equals

[1] We need to draw an explicit distinction between dP_i and the predictive densities p_i introduced earlier, since they are actually related to different dominating measures.

$$P(\delta = 0|x) = \frac{a dP_0}{a dP_0 + (1 - a) dP_1}. \qquad (14)$$

We consider the case where F^θ is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^p$, with density function $f(x|\theta)$. The corresponding predictive densities on M_0 for a single observation x_i and for the sample x are given by

$$p_0(x_i) = \int_\theta f(x_i|\theta)\pi_0(\theta)d\theta \qquad (15)$$

and

$$p_0(x) = \int_\Theta \prod_{i=1} f(x_i|\theta)\pi_0(\theta)d\theta; \qquad (16)$$

respectively. The posterior probability of model M_0 under a predictive coherent prior Dirichlet measure is then characterized in the following theorem.

Theorem 3.1 : If F^θ is absolutely continuous w.r.t. the Lebesgue measure with density $f(x|\theta)$ and if the Dirac prior on M_1 is predictive-coherent in the sense of (12), then the posterior probability of M_0 is given by

$$P(\delta = 0|x) = \frac{a p_0(x)}{a P_0(x) + (1 - a)\alpha_n \prod_{i=1}^n p_0(x_i)} \mathbb{1}_{\{d_n = n\}}; \qquad (17)$$

where $a = P(\delta = 0); p_0(x_i)$ and $p_0(x)$ are defined in (15) and (16), respectively, d_n denotes the number of distinct values in the sample x and α_n the predictive probability of no ties on M_1, which is given by

$$\alpha_n = P_1(d_n = n) = \frac{\Gamma(n_a) \cdot n_a^n}{\Gamma(n_a + n)} \qquad (18)$$

$\square$

Proof : The key issue is that of choosing a suitable dominating measure. Let L denote the Lebesgue measure on $\mathbb{R}^p$. Clearly, $dP_0/dL = p_0$. Since P_1, as defined in (11), can generate ties, we decompose it as follows

$$P_i = \alpha_n P_{11} + (1 - \alpha_n) P_{12};$$

where $P_{11} = P_1(\;|d_n = n)$ and $P_{12} = P_1(\;|d_n \neq n)$. We note that P_{11} and P_{12} are mutually singular since, in particular, $P_{11}(d_n = n) = 1$ and $P_{12}(d_n = n) = 0$. Note furthermore that $P_0(d_n = n) = 1$ by the absolute continuity of F^θ. Let choose $P_* = L + P_{12}$ as the dominating measure. It immediately follows that $dP_0/dP_* = p_0 \mathbb{1}_{\{d_n=n\}}$. Furthermore, it follows from (11) that, conditionally on $d_n = n$, the x_i' s are i.i.d. F_a on M_1, whence

$$\frac{dP_1}{dP_*} = \alpha_n \prod_{i=1} p_0(x_i) \mathbb{1}_{\{d_n=n\}} + (1 - \alpha_n) \mathbb{1}_{\{d_n<n\}}$$

and (17) follows. It also follows from (11) that $\alpha_1 = 1$ and $\alpha_{i+1} = \frac{n_a}{n_a+i} \alpha_i$ which implies (18) and completes the proof. ∎

If M_0 consists of a simple hypothesis, i.e. if $F^\theta \equiv F_a$, then $p_0(x) = \prod_{i=1}^{n} p_0(x_i)$ and the following corollary obtains.

Corollary 3.2 : Under the conditions of theorem 1 and if $F^\theta \equiv F_a$ on M_0, then

$$P(\delta = 0|x) = \frac{a}{a + \alpha_n(1 - a)} \mathbb{1}_{\{d_n=n\}}. \tag{19}$$

∎

It is important to note that the predictive density P_1, as defined in (11), is degenerate. It follows that Bayesian model selection procedures based upon posterior odds fundamentally amount to a test of continuity (M_0) versus discreteness (M_1), a conclusion which is fully supported by the behavior of the posterior probabilities in (17) and (19). Note, in particular, that $P(\delta = 0|x) \equiv 0$ as soon as $d_n < n$. M_0 is rejected as soon as a tie is observed. Conversely, if d_n stays equal to n, then $\alpha_n \to 0$ and $P(\delta = 0|x) \to 1$ as $n \to \infty$, independently of the actual "validity" of M_0. In the absence of ties, M_0 ends being always accepted for large enough sample sizes. The same holds for any sample size if $n_a = 0$ (non-informative prior on M_1). In other words, the "implicit null hypothesis" of a posterior odds "test" procedure consists of the set of all continuous distributions, despite the fact that the Dirichlet prior on M_1 is dense ! This "extreme" behavior of posterior odds unequivocally suggest that they are fundamentally inadequate as an instrument for testing the "validity" of M_0 by confronting it to a general nonparametric alternative M_1.

We conclude this discussion with two additional comments. Firstly, the behavior of the posterior odds is fundamentally unaffected if we remove the assumption of predictive-coherency. It can be proved that if F_a no longer satisfies condition (12), the posterior probabilities in (17) and (19) nevertheless remain valid, except that $p_0(x_i)$ is to be replaced by $p_1(x_i)$, the predictive density of x_i on M_1. Secondly, it is not possible to eliminate the intrinsic "bumpy" character of the M_1-predictive distribution by increasing the prior weight attached to F_a, i.e. by increasing n_a along with n, the actual sample size. It is well known that $d_n(n_a \ln n)^{i\,1}$ converges almost surely to 1 on M_1 as n increases -see e.g. Rolin (1993). One would, therefore, be tempted to set n_a equal to $n{=}\ln n$, a choice which preserves the consistency of the posterior distribution since $n_a{=}n$ still tends to zero as n tend to infinity. This choice, however, does not prevent $®_n$ from (rapidly) tending to zero as n tends to infinity. Actually, it can be shown that, in order to have $®_n$ tending toward a constant, we have to choose n_a's of the order of n^2. Such choices are clearly unacceptable since, in particular, they destroy the consistency of posterior distributions on M_1.

4 Encompassing

We first briefly present the encompassing principle per se, using notation introduced earlier, and then apply it to the pair of models under consideration.

4.1 General principle

As already discussed in the introduction our objective is not that of choosing between M_0 and M_1 but rather that of (in)validating M_0 by analyzing whether results derived on M_1 can be reproduced within M_0. Within a Bayesian framework, results to be encompassed typically take the form of a posterior density on a functional ¸ of the "parameter" F associated with M_1. Drawing upon the general presentation in Florens et al (1997), we briefly discuss here how the encompassing principle applies to the models introduced in section 2.

The key step consists of prolonging M_0 into a model $M_¤$ which is characterized by a joint probability on $(µ; X)$ and F. This extension is obtained by adding onto M_0 a conditional distribution for F, given µ. The joint distribution of $(F; µ; x)$ on $M_¤$ is then characterized by the density

$$¼(F; µ; x) = ±(F jµ):[p_0(xjµ):¹_0(µ)]: \qquad (20)$$

Note that ¼ emboddies the key assumption that F and x are mutually independent, conditionally on µ, an assumption we shall qualify further in a moment. The "M_0-posterior density" of F is then defined as the posterior density of F which obtains from π. It can be rewritten as

$$\gamma_0(F|x) = \int_\Theta \delta(F|\theta)\mu_0(\theta|x)d\theta; \qquad (21)$$

from which we can obtain an M_0-posterior density for λ say, $\gamma_0(\lambda|x)$. Encompassing compares the two posterior densities for λ : $\gamma_1(\lambda|x)$, as obtained on M_1 and $\gamma_0(\lambda|x)$ which constitutes our "reinterpretation" of $\gamma_1(\lambda|x)$ within the context of the enlarged M_0.

There are two key reasons for imposing independence between F and x, conditionally on θ. Firstly, at a heuristic level, encompassing aims at reinterpreting M_1-posterior densities within M_0, without additional data processing beyond that already incorporated in the posterior density $\mu_0(\theta|x)$. Furthermore, from the viewpoint of M_0, it appears natural to assume that θ is a sufficient parametrization of M_0, i.e. that $p_0(x|\theta; F) \equiv p_0(x|\theta)$, an assumption which immediately validates (20).

Secondly, at a more formal level, if the transition δ were allowed to depend on x, then it would always be possible to produce a perfect match between $\gamma_1(\lambda|x)$ and $\gamma_0(\lambda|x)$, immediately voiding the encompassing principle from any meaning. As discussed in Florens et al (1997), the assumed conditional independence between F and x enables us to formally reinterpret encompassing as a concept of "sufficiency among models", dual to the traditional concept of sufficiency (among statistics), as defined by Blackwell (1951, 1953) or Lecam (1964). That concept provides an unambiguous statistical foundation to the commonly observed scientific pratice of verifying that new theories or models be capable of accounting for findings ("failures" as well as "successes") obtained from earlier models.

There remains to discuss how to formalize comparisons between γ_1 and γ_0. An operational procedure consists of computing a measure of "divergence" between γ_1 and γ_0, say, $\varepsilon_\lambda(x) = D(\gamma_1(\lambda|x); \gamma_0(\lambda|x))$. Such a measure, which depends upon the observed sample x, is refered to as to a measure of the specificity of M_1 relative to M_0 (in reference to λ). As discussed e.g. in Florens et al. (1997) different choices are available for the divergence D. In the present paper, we restrict our attention to the entropy measure

$$\varepsilon_\lambda(x) = \int_\Lambda \ln\left[\frac{\gamma_0(\lambda|x)}{\gamma_1(\lambda|x)}\right]\gamma_0(\lambda|x)d\lambda; \qquad (22)$$

which proves to be particularly operational for the problem under consideration.

Two additional issues need to be addressed before we close this brief presentation of encompassing. Firstly and foremost, we have to select an appropriate transition density ±(Fjµ), often refered to as to a "Bayesian Pseudo-True Value" (hereafter BPTV). In the present paper, we use a BPTV which is defined as the M_0-(sample) expectation of the M_1-posterior density of F, that is to say

$$\pm(F j\mu) = \int {}^{o}_{1}(F j\tilde{x}) p_0(\tilde{x} j\mu) d\tilde{x}: \qquad (23)$$

This transition offers two keys advantages : (i) As we shall illustrate below, it can be evaluated at a high level of generality under a combination of Monte Carlo simulation and kernel smoothing; and (ii), it produces consistent encompassing test procedures since, under general technical conditions, it converges toward classical pseudo-true values (which are formally defined as plims on M_0 of Maximum Likelihood estimators).

The sample size $\tilde{n}$ in (23) need not be set equal to n, the actual sample size. For the ease of reference, we shall explicitely distinguish between n and $\tilde{n}$ in all subsequent formulae. This being said, there are good reasons for setting $\tilde{n}$ = n. Foremost, it is common pratice to condition BPTV's upon exogenous variables and in the context consideration, sample size truly is exogenous. Also, though we do not specifically discuss here asymptotic encompassing - see e.g. Florens and Richard (1998) for asymptotic analysis in the context of finite parameter spaces - consistency of encompassing test procedures under BPTV's does require that n and $\tilde{n}$ both tend to infinity at a common rate. Finally, setting $\tilde{n}$ = n facilitates calibration. We note also that, had we set $\tilde{n}$ = 1, the BPTV would collapse into a (degenerate) classical pseudo-true value.

A second problem is that of interpreting the actual value obtained for $^{2}_{¸}(x)$. An obvious and fairly simple "calibration" procedure consists of evaluating (by Monte Carlo simulation) its distribution on the M_0-predictive distribution of x and of computing the probability of drawing a value larger than that which actually obtained (such a probability can usefully be reinterpreted as a Bayesian encompassing p-value). This "exact" calibration procedure can also usefully contribute "standardizing" specificity measures that would be evaluated under alternative divergence measures.

Finally, we ought to mention that more "decision oriented" Bayesian encompassing procedures could usefully be considered under specific circumstances. If, for example, the functionals ¸ happened to be of interest to the proprietor of M_0 within a given decisional context (or if it was thought to be desirable that decisions reached on M_1 be encompassed within M_0), then differences in posterior expected losses would provide a natural metric to evaluate the specificity of M_1 relative to M_0. Assume, for example, that M_1 were paired with a decision $a \in A$ and a loss function $l(a;\lambda)$. One would then compute the following posterior expected losses

$$\begin{aligned} l^*_1(x) &= \min_{d \in D} \int_{\Lambda} l(d;\lambda)\gamma_1(\lambda|x)d\lambda; \\ l^*_0(x) &= \min_{d \in D} \int_{\Lambda} l(d;\lambda)\gamma_0(\lambda|x)d\lambda \\ &= \min_{d \in D} \int_{\Lambda\times\Theta} l(d;\lambda)\mu_0(\theta|x)\delta(\lambda|\theta)d\lambda d\theta: \end{aligned} \tag{24}$$

The difference between $l^*_0(x)$ and $l^*_1(x)$ would provide an obvious measure of how well M_0 encompasses M_1 for the specific decision problem under consideration. Here again, that difference could be calibrated under the M_0-predictive density of x.

We conclude this presentation of encompassing by mentioning here that the use of measures of divergence (entropy,...) has often been advocated as a simple alternative to a full decision-oriented evaluation. See e.g. DeGroot (1970)

4.2 Nonparametric encompassing

We now apply the encompassing procedure we just described to the pair of models introduced in section 2. We have already established that the M_1-posterior distribution of F is $Di(n_p; F_p)$ where n_p and F_p are defined in (9). It follows that the BPTV introduced in (23) is a mixture of Dirichlet processes. Using the density notation introduced earlier we can rewrite the M_0-posterior density of F as follows :

$$\begin{aligned} \gamma_0(F|x) &= \int_{\Theta} \delta(F|\theta)\mu_0(\theta|x)d\theta \\ &= \int_{\Theta}\int_{\mathbb{R}^{n_p}} \gamma_1(F|\tilde{x})p_0(\tilde{x}|\theta)\mu_0(\theta|x)d\tilde{x}d\theta \\ &= \int_{\mathbb{R}^{n_p}} \gamma_1(F|\tilde{x})p_0(\tilde{x}|x)d\tilde{x}; \end{aligned} \tag{25}$$

where

$$p_0(\tilde{x}|x) = \int_{\Theta} p_0(\tilde{x}|\theta)\mu_0(\theta|x)d\theta \tag{26}$$

denotes the M_0-posterior predictive distribution of $\tilde{x}$, given x. It follows, therefore, that $\gamma_0(F|x)$ is itself a mixture of Dirichlet measures. Actually, the only difference between $\delta(F|x)$ and $\gamma_0(F|x)$ lies in the mixing distribution which is $p_0(\tilde{x}|\theta)$ for δ and $p_0(\tilde{x}|x)$ for γ_0. M_1 and M_0-expectations of F are given by

$$\begin{aligned}
E_1(F) &= F_a, \\
E_1(F|x) &= \frac{n_a}{n_a+n}F_a + \frac{n}{n_a+n}F_n, \\
E_0(F|\theta) &= \frac{n_a}{n_a+\tilde{n}}F_a + \frac{\tilde{n}}{n_a+\tilde{n}}F^{\theta}, \\
E_0(F|x) &= \frac{n_a}{n_a+\tilde{n}}F_a + \frac{\tilde{n}}{n_a+\tilde{n}}E_0(F^{\theta}|x), \qquad (27)
\end{aligned}$$

respectively. We note that, under the predictive coherent prior in (12), we have $E_0(F) = E_{\theta}[E_0(F|\theta)] = F_a = E_1(F)$. Furthermore, the comparison between the M_0-and M_1-posterior expectations of F amounts to a comparison between $E_0(F^{\theta}|x)$, the M_0 posterior expectation of F^{θ} and F_n, the empirical distribution of x.

We first demonstrate that analytical characterizations of the transition $\delta(F|\theta)$ and of the M_1-posterior distribution $\gamma_1(F|x)$ are intractable. According to Ferguson (1973), it suffices to derive the distribution of the vector $\{F(B_l)\ l : 1 \to L\}$ for any non trivial partition $\{B_l, l : 1 \to L\}$ of $\mathbb{R}^p$, notwithstanding the fact that there might be specific partition(s) of interest to be encompassed by M_0 (The latter would be included in the definition of the functional "of interest" $\lambda = \varphi(F)$ which is introduced below).

Relative to an arbitrary partition $\{B_l, l : 1 \to L\}$, a sufficient statistic associated with the sample x is given by $\{n_l, l : 1 \to L\}$, where n_l denotes the number of x_i's in B_l. Therefore, the M_1-posterior density of $\{F(B_l), l : 1 \to L\}$ is given by

$$\mu_1(F(B_1)),\ldots,F(B_L)|x) = \Gamma(n_a+n)\prod_{l=1}\frac{(F(B_l))^{n_aF_a(B_l)+n_l-1}}{\Gamma(n_aF_a(B_l)+n_l)}. \qquad (28)$$

A similar expression applies to the M_1-posterior distribution of $\{F(B_l), l : 1 \to L\}$ conditional on $\tilde{x}$, except that the n_l's in (28) are replaced by $\tilde{n}_l$'s. Since, furthermore, the $\tilde{x}_i$'s are i.i.d. F^{θ} on M_0, the $\tilde{n}_l$'s follow a multinomial distribution with parameters $\tilde{n}$ and $\{F^{\theta}(B_l)\ , l : 1 \to L\}$. It immediately follows that $\delta(F(B_1),\ldots,F(B_L)|\theta)$ is given by

$$\pm(F(B_1);::::;F(B_L)j\mu) = \mathbf{P}_{f\tilde{n}_l g2S_{L\,i\,1}} \mathbf{h}\; i\,(n_a+\tilde{n})\; \mathbf{Q}^{L}_{l=1} \frac{(F(B_l))^{n_aF_a(B_l)+\tilde{n}_l\,i\,1}}{i\,(n_aF_a(B_l)+\tilde{n}_l)} \mathbf{i} : \mathbf{h}\; \tilde{n}\; \mathbf{Q}^{L}_{l=1} \frac{(F^{\mu}(B_l))^{\tilde{n}_l}}{\tilde{n}_l!} \mathbf{i} ; \tag{29}$$

where $S_{L\,i\,1} = f(\tilde{n}_1;::::;\tilde{n}_L);\tilde{n}_l\ 2\ \mathbb{N};\ \mathbf{P}^{L}_{l=1}\tilde{n}_l = \tilde{n}g$. Such summations have no (simple) analytical solutions and, furthermore, include very large numbers of terms for all but small $\tilde{n}$ and/or L. It also follows that $^o{}_1(Fjx)$ is itself analytically intractable.

On the other hand, we can easily simulate trajectories of F under its M_0- and M_1-posterior distributions by relying upon powerful representations of the Dirichlet process, as found, e.g., in Sethuraman (1994) or Rolin (1993). Let F_i^x denote a realisation of the M_i-posterior distibution of $F(i = 1;2)$. The following representation applies to F_1^x

$$F_1^x = (1\ i\ ^\circ) \sum_{k=1}^{\mathbf{X}} ®_k\pm_{»_k} + ^\circ \sum_{i=1}^{\mathbf{X}} {}^{-}{}_i\pm_{x_i}; \tag{30}$$

where

(i) $^\circ ?? f®_k;k:1\ !\ 1g??f»_k;k:1\ !\ 1g??f^{-}{}_i;i:1\ !\ ng;$

(ii) $^\circ\ »\ B(n;n_a)$, where $B(a;b)$ denotes the Beta distribution with parameters $a > 0$ and $b > 0$;

(iii) The $»_k$'s are i.i.d. F_a;

(iv) $®_k = v_k\ \mathbf{Q}^{k\,i\,1}_{l=1}(1\ i\ v_l)$ for $k:1\ !\ 1$, where the v_l's are i.i.d. $B(1;n_a)$;

(v) The $^{-}{}_i$'s are uniformly distributed on the simplex $S_{n\,i\,1}$, i.e. $(^{-}{}_1;::::;^{-}{}_n)\ »\ Di(n;(1:::1))$;

and where, as above, $\pm_{x_i}(\pm_{»_k})$ denotes the Dirac measure at point $x_i(»_k)$, i.e., $\pm_{x_i}(B) = 1\!\!1_B(x_i)$.

A similar representation applies to F_0^x, except that x in (30) is replaced by $\tilde{x}$ which, according to (26), is generated as follows : μ is drawn from $^1{}_0(\mu jx)$ and, conditionally on μ, the $\tilde{x}_i$'s are drawn independently from one another from F^{μ}.

Using these representations, we can generate repeated draws from the posterior distributions $^o{}_0$ and $^o{}_1$. The next issue to be addressed is that of using there draws to evaluate a divergence measure between ν_1 and ν_0. It is important to immediately point out that ν_1 and ν_0 are mutually singular. For example, F_1^x in (30) assigns a non zero probability to x_1, that is to say, the set of measures on $(\mathbb{R}^p; \mathcal{B}_p)$ which assign positive probability to x_1 has measure one on $\nu_1(F|x)$. On the other hand F_0^x assigns zero probability to x_1, since the probability that $\tilde{x}_1 = x_1$ on $p_0(\tilde{x}|x)$ is zero. It follows that conventional divergence measures will automatically be maximal, leading to automatic rejection of the corresponding encompassing test procedure (a problem which is closely related to that already discussed in section 3 for posterior odds). We should then consider metrics which can produce non degenerated comparisons of mutually singular distributions. Skorohod's distances might provide a conceptual solution in that respect but it is far from being obvious that they could lead to operational procedures.

A fully operational solution consists of restricting the encompassing comparison to the M_0- and M_1-posterior distributions of an appropriate functional of F_1 say

$$\lambda = \varphi(F) \in \mathbb{R}^k; \qquad (31)$$

where λ would represent "parameters of interest" within M_1 which would, therefore, be obvious targets for an encompassing test. The key advantage offered by this approach lies in the finding that, under very mild technical condition, the M_0- and M_1-posterior distributions of λ are absolutely continuous w.r.t. the Lebesgue measure -see e.g. Florens and Rolin (1994) - whence divergence comparisons do apply to λ. Typical choices for λ are :

(i) $\lambda = (F(B_l); l : 1 \to l)$ for a measurable partition $(B_1; \ldots B_L)$ of $\mathbb{R}^p$ which regroups "regions of interest" in the sample space ; (ii) $\lambda = \int_{\mathbb{R}^p} h(x)F(dx)$, where h denotes a Borel function from $\mathbb{R}^p$ to $\mathbb{R}^k$. For example a comparison of first and second order moments would obtain for $\lambda = (x; xx')$ and $k = p + \frac{1}{2}p(p+1)$.

The simulation method described above is then exploited in the following way to produce an operational encompassing test procedure. A set of randomly drawn trajectories $\{(\tilde{F}_{1;r}^x; \tilde{F}_{0;r}^x); r : 1 \to R\}$ is transformed into a corresponding set of random drawns for λ, say $\{\tilde{\lambda}_{1;r}^x; \tilde{\lambda}_{0;r}^x); r : 1 \to R\}$. Kernel estimates of the posterior densities $\nu_1(\lambda|x)$ and $\nu_0(\lambda|x)$ are then evaluated, which are denoted by $\hat{\nu}_1(\lambda|x)$ and $\hat{\nu}_0(\lambda|x)$, respectively. A measure of the λ-specificity of M_1 relative to M_0 is then given by

$$\hat{\lambda}(x) = \frac{1}{R} \sum_{r=1} \ln \left[\frac{\hat{\nu}_0(\tilde{\lambda}_{0;r}^x|x)}{\hat{\nu}_1(\tilde{\lambda}_{0;r}^x|x)} \right]; \qquad (32)$$

Finally, in order to calibrate $ª_¸(x)$ on M_0, we generate S (i.i.d.) samples from $p_0(x)$, the M_0-predictive distribution of x and compute $ª_¸(x)$ for each such sample, a procedure which produces S i.i.d. draws from $ª_¸(x)$ on p_0, from which a Bayesian p-value immediately obtains as the proportion of draws which exceeds the actual sample value $ª_¸(x)$.

5 An application : validation of an exponential distribution

Let X denote a positive scalar random variable, such as a lifetime. Our application is based upon a sample x of size n = 50 which we drew from an exponential distribution with parameter $µ = 1$. Let M_0 consists of an exponential distribution with parameter $µ > 0$ under a natural conjugate gamma prior density for µ with parameters $a_0 > 0$ and $º_0 > 0$. The corresponding sample and prior density functions are given by

$$p_0(xjµ) = µ^n e^{iµt}; \text{ and} \tag{33}$$

$$¹_0(µ) = \frac{a_0^{º_0}}{i(º_0)} µ^{º_0 i 1} e^{i a_0 µ}; \tag{34}$$

respectively, where $t = \mathbf{P}_{i=1}^{n} x_i$ is a sufficient statistic. Actual calculations are based upon the values $a_0 = º_0 = 2$. The M_0-posterior distribution is itself a gamma distribution with parameters $a_¤ = a_0 + t$ and $º_¤ = º_0 + n$. The M_0-predictive densities in (15) and (16) are given by

$$p_0(x_i) = º_0 a_0^{º_0}(a_0 + x_i)^{i(º_0+1)}; \text{ and} \tag{35}$$

$$p_0(t) = \frac{i(º_0 + n)}{i(º_0)} a_0^{º_0}(a_0 + t)^{i(º_0+n)}; \tag{36}$$

respectively. We note in passing that $p_0(x_i)$ represents a Pareto distribution.

On M_1, we use a non-informative Dirichlet prior measure on F by setting $n_a = 0$. In view of the "weakly" informative M_0-prior ($a_0 = º_0 = 2$), imposing the predictive-coherency condition (12) makes little difference while setting $n_a = 0$ simplifies the representation of F_1^x in (30), which is now given by

$$F_1^x = \sum_{i=1} \beta_i \delta_{x_i}; \tag{37}$$

where $\{\beta_i; i : 1 \to n\} \sim Di(n; (1; \ldots; 1))$. A similar expression applies to F_0^x with x being replaced by $\tilde{x}$ to be drawn from the M_0-predictive posterior distribution in (26).

Finally, implicitely restricting M_1 to distributions with finite first and second order moments, we consider here that the latter are the parameters of interest to be encompassed by M_0, i.e.

$$\theta_i = \int_0^{\infty} x^i F(dx); \text{ for } i = 1; 2: \tag{38}$$

Following the procedure described in section 4.2, we draw two samples of size R from $\pi_1(\theta|x)$ and $\pi_0(\theta|x)$, respectively, that is to say

$$\tilde{\theta}^x_{1;r} = \sum_{j=1} \tilde{\beta}_{1;j;r} \left(\begin{array}{c} x_j \\ x_j^2 \end{array} \right); \text{ and } \tilde{\theta}^x_{0;r} = \sum_{j=1} \tilde{\beta}_{0;j;r} \left(\begin{array}{c} \tilde{x}_{j;r} \\ \tilde{x}_{j;r}^2 \end{array} \right); \tag{39}$$

where $\{\tilde{\beta}_{1;j;r}; j : 1 \to n\}$ and $\{\tilde{\beta}_{0;j;r}; j : 1 \to n\}$ for $r : 1 \to R$ are i.i.d. drawns from a $Di(n; (1; \ldots; 1))$ distribution, and $\{\tilde{x}_{j;r}; j : 1 \to n\}$ for $r : 1 \to R$ are i.i.d. draws from $p_0(\tilde{x}|x)$ (in pratice, $\tilde{x}_r$ is drawn from (33), conditionally on $\mu = \tilde{\mu}_r$, where $\tilde{\mu}_r$ itself is drawn from (34)). Next, we compute bivariate kernel estimates for the posterior distributions of θ say

$$\hat{\pi}_i(\theta|x) = \frac{1}{R} \sum_{r=1} K_2(\theta - \tilde{\theta}^x_{i;r}); \text{ for } i = 1; 2; \tag{40}$$

where K_2 denotes the kernel

$$K_2(u_1; u_2) = \prod_{i=1} \frac{1}{h_{i;R}} \phi \left(\frac{u_i}{h_{i;R}} \right); \tag{41}$$

ϕ denotes the standardized normal density and $h_{i;R} = S_i R^{-1/5}$, where S_i denotes the estimated standard deviation of the $\tilde{\theta}^x_{i;r}$ for $i = 1; 2$. The θ-specificity of M_1 relative to M_0 then obtains by application of (32) and equals 0.58. In the present application we use a Monte Carlo sample size $R = 1; 000$.

We generated 1,000 auxiliary samples of size $n = 50$ from the M_0-predictive density $p_0(t)$ in (36). Actually, each such draw is generating by drawing first μ from the prior (34) and then x conditionally on μ from (33). The computations we just described are applied to each auxiliary sample in turn in order to produce 1,000 draws from the M_0-predictive distribution of $\hat{\xi}(x)$. An histogram of the latter is reproduced in figure 1 and a few selected "critical values" are reported in table 1. The M_0-predictive mean and standard deviation of $\hat{\xi}(x)$ equal 0.8104 and 0.3374, respectively.

Finally, we also evaluated the ξ-specificities of M_1 relative to a variety of "invalid" M_0's, all constrained to have expectations equal to 1. The entropy results are reported in table 2, together with the corresponding Bayesian p-values which obtain from the histogram in figure 1. In contrast with the valid exponential model, these invalid models clearly fail to encompass M_1 relative to ξ.

6 Conclusion

We have proposed a flexible and fully operational Bayesian procedure for examining whether or not a tentative model M_0 encompasses results derived under a general nonparametric model M_1. It is important to emphasize the fact that encompassing is not a model selection procedure since, in particular, M_1 generally is not meant to be interpreted as an alternative to M_0 but rather as an instrument in testing the (in)validity of M_0. Clearly, broad failure to encompass characteristics of interest of M_1 would lead to the conclusion that M_0 is seriously deficient. Whether such deficiencies would lead one to try to refine M_0 further or instead to trash it can only be answered in the context of the specific application under consideration after careful examination of the said deficiencies.

Our method offers considerable flexibility in this respect relative, in particular, to the selection of results to be encompassed which would typically depend upon the objectives of the proprietor of M_0. Encompassing also requires the selection of a transition probability and of a divergence measure. Though such choices introduce a degree of arbitrariness in our procedure, the pair consisting of the entropy measure (22) and of the BPTV (23) provides a natural Bayesian extension of concepts widely used in statistics. Furthermore, arbitrariness is tempered by the fact that all encompassing results are M_0 calibrated in the end. As briefly discussed in the paper, much arbitrariness can also be removed by adopting a more decision oriented measure of encompassing differences.

In view of the inherent complexity of the task of constructing empirical econometric models, we strongly believe that encompassing can provide an invaluable tool for carefully investigating potential deficiencies of a model and suggesting avenues for further improvements.

Figure 1 :
M_0-predictive distribution of ¸-specificity of M_1.

Table 1. Estimated critical values of the ¸-speci¯city

1 ¡ ®	0,90	0,91	0,92	0,93	0,94	0,95	0,96	0,97	0,98	0,99
$C_{1\ ¡\ ®}$	1,28	1,30	1,33	1,35	1,39	1,44	1,49	1,59	1,70	1,94

Table 2. Bayesian p-values

M_0 distribution	Variance	Entropy	Prob-value
Exponential (1)	1; 000	0,58	0,829
Weibull (2)	0; 273	3,17	0,000
Gamma ($\frac{1}{2}; \frac{1}{2}$)	2,000	1,43	0,052
Gamma (2,2)	0; 500	1,69	0,022
Gamma (4,4)	0; 250	1,95	0,009
Uniform (0,2)	0; 333	1,83	0,015
Lognormal (¡ $\frac{1}{4}; \frac{1}{2}$)	0; 649	1,31	0,086
Lognormal (¡ $\frac{1}{2}$¾^2; ¾^2) ¾^2 = 0; 2231	0; 250	2,88	0,000

References

Antoniak C.E. (1974) "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems" Annals of Statistics,2, 1152-1174.

Berk, R. (1966) "Limiting behavior of posterior distributions when the model is incorrect", The Annals of Mathematical Statistics, 37, 51-58.

Blackwell D. (1951) "Comparison of experiments", Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probabilities, Berkeley : University of California Press, 93-102.

Blackwell D. (1953) "Equivalent comparison of experiments", The Annals of Mathematical Statistics, 24, 265-272.

DeGroot M.H. (1970) "Optimal statistical decisions", McGraw-Hill, Inc, New-York.

Escobar, M.D. and M. West (1995) "Bayesian density estimation and inference using mixtures", Journal of the American Statistical Association, 90, 577-588.

Ferguson T.S. (1973) "A Bayesian analysis of some non parametric problems", Annals of Statistics, 1, 209-230.

Ferguson T.S. (1974) "Prior distributions on spaces of probability measures", Annals of Statistics, 2, 615-629.

Florens J.P. and M. Mouchart (1993) "Bayesian testing and testing Bayesians", Handbook of Statistics, vol.11, 65, 303-334, Maddala G.S., C.R. Rao and H.D. Vinod, Edts, Elsevier Science Publishers, .

Florens J.P. and J.M. Rolin (1994) "Bayes, Bootstrap, Moments" Institut de Statistique D.P. 9413, Universit¶e Catholique de Louvain, Belgium.

Florens J.P., D.P. Hendry and J.F. Richard (1996) "Encompassing and specificity", Econometric Theory, 12, 620-656.

Florens J.P., C. Protopopescu and J.F. Richard (1997) "Identification and estimation of a class of game theoretic models", University of Toulouse.

Florens J.P. and J.F. Richard (1998) "Encompassing in finite parametric spaces", University of Toulouse.

Huber P.J. (1967) "The behaviour of maximum likelihood estimates under non standard conditions" in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol.1, 105-123, L. Lecam and J. Neyman, Edts, University of California Press, .

Lavine, M (1992) "Some aspects of polya tree distributions for statistical modelling", The Annals of Statistics, 20, 1222-1235.

Lavine, M. (1994) "More aspects of Polya tree for statistical modelling", The Annals of Statistics, 22, 1161-1176.

LeCam, L. (1964), "Sufficiency and approximate sufficiency", The Annals of Mathematical Statistics, 35, 1419-1455.

Lenk, P.J. (1988) "The logistic normal distribution for Bayesian, nonparametric, predictive densities", Journal of the American Statistical Association, 83, 509-516.

Lenk, P.J. (1991) "Towards a practicable bayesian nonparametric density estimator", Biometrika, 78, 531-543.

Leonard, T. (1978) "Density estimation, stochastic processes and prior information", The Journal of Royal Statistical Society, B 40, 113-146

Lo, A.Y. (1984) "On a class of Bayesian nonparameteric estimates I density estimates", The Annals of Statistics, 12, 351-357.

Marshall, R., M. Meurer, J.F. Richard and W. Stromquiest (1994) "Numerical analysis of asymmetric first-price auctions", Game and Economic Behavior, 7, 193-220.

Raiffa, H. and R. Schlaifer (1961), Applied statistical decision theory, Boston : Division of Research, Harvard Business School.

Révész P. (1984) "Density estimation" in Handbook of Mathematical Statistics vol. 4, 531-549, P.R. Krishnaiah and P.K. Sen, Edts, Elsevier Science Publishers, .

Rolin J.M. (1992) "Some useful properties of the Dirichlet process", Institut de Statistique D.P. 9202, Université Catholique de Louvain, Belgium.

Rolin J.M. (1993) "On the distribution of jumps of the Dirichlet process", Institut de Statistique D.P. 9302, Université Catholique de Louvain, Belgium.

Sethuraman J. (1994) "A constructive definition of Dirichlet priors", Statistica Sinica, Vol 4(2), 639-650.

White H. (1982) "Maximum Likelihood estimation of misspecified models", Econometrica, 50, 1-26.