Encompassing and Specificity

J.P. FLORENS
GREMAQ and IDEI, Toulouse University, France (33 6215 0606)
David F. HENDRY
Nuffield College, Oxford, UK (44 865 278554)
and
Jean-François RICHARD[*]
University of Pittsburgh, USA (1 412 648 1750).

[*] This research was financed in part by grants R000233447 and R000234954 from the UK Economic and Social Research Council. Helpful comments from Peter Phillips and three anonymous referees are gratefully acknowledged.

Running head:

Encompassing and Specificity

Address for Proofs:

David F. Hendry,
Nuffield College,
Oxford, OX1 1NF,
UK.

ABSTRACT

A model $\mathcal{M}$ is said to encompass another model $\mathcal{N}$ if the former can explain the results obtained by the latter. In this paper we propose a general notion of encompassing which covers both classical and Bayesian viewpoints and essentially represents a concept of sufficiency among models. We introduce the parent notion of specificity which aims at measuring lack of encompassing. Tests for encompassing are discussed and the test statistics are compared to Bayesian posterior odds. Operational approximations are offered to cover situations where exact solutions cannot be obtained.

RESUME

Un modèle $\mathcal{M}$ enveloppe un modèle $\mathcal{N}$ si les résultats obtenus par le second modèle peuvent être expliqués par le premier. Dans cet article, nous proposons une notion générale d'enveloppement essentiellement considérée comme une propriété d'exhausitivité entre modèles. Nous introduisons alors la notion de spécificité comme mesure du défaut d'enveloppement. Des tests d'enveloppement sont présentés et comparés aux procédures de choix de modèles fondées sur les probabilités a postériori de modèles ('Posterior odds'). Des approximations opérationnelles sont enfin proposées pour analyser des situations dans lesquelles les solutions exactes ne peuvent pas être obtenues.

1 Introduction

One 'model' $\mathcal{M}$ is said to encompass another 'model' $\mathcal{N}$ if the former can account for the results obtained by the latter. This notion has long been accepted as a critical component of research strategies in most sciences. Numerous applications in the econometric literature include investigations of the implications of each of a series of models for the others. Recent developments in econometrics have opened the way to formalizations of the notion of encompassing (see *inter alia* Hendry and Richard [20], [22], Mizon [27], Mizon and Richard [28], Florens and Mouchart [11], [12], Florens, Mouchart and Rolin [13], and Govaerts, Hendry and Richard [19]. The object of this paper is to build upon these earlier contributions and to propose a rigorous and general definition of encompassing, which can accommodate classical and Bayesian viewpoints, parametric, semi-parametric and non-parametric procedures and which, in line with recent econometric developments, does not require the models under consideration to be correctly specified.

Formal definitions are offered below, but a brief heuristic discussion helps set the scene for our analysis. First, we distinguish between the data generating process (DGP) and an inference procedure (IP). The DGP is the actual mechanism, conceptualized as a class of sampling probabilities $\mathcal{P} = \left\{P^{\theta}, \boldsymbol{\theta} \in \Theta\right\}$ on a measurable sample space $(S, \mathcal{S})$. Θ is a set of 'parameters' (possibly functional ones) indexing $\mathcal{P}$ but the analysis does not require the DGP to be specified in full. For example, $\mathcal{P}$ might consist of the set of all independent identically distributed (iid) probability measures admitting a preassigned number of moments. IPs are procedures which are designed to draw inferences on functions of $\boldsymbol{\theta}$ valued in a set A (which is typically of lower dimensionality than Θ itself). Examples are estimators, i.e. functions from S into A, and posterior distributions, i.e. probability measures on $(A, \mathcal{A})$ conditional on the elements of S. IPs may be associated with a maximization criterion (maximum likelihood or generalized method of moments) or follow from the application of Bayes theorem to an auxiliary sampling model, which is typically 'mis-specified' relative to the DGP.

Encompassing is reinterpreted as a concept of sufficiency between IPs, 'dual' to that of sufficiency among sampling processes. An IP M encompasses another IP N if the results derived from N can be reproduced within M without requiring further processing of sample information, beyond that already associated with M, i.e. if the results of N are 'contained' in those obtained from M.

We introduce the concept of exact encompassing applicable to finite sample situations. A procedure M from S to A exactly encompasses a procedure N from S to B if there exists a pseudo-true value Δ from A to B such that $N = \Delta \circ M$. Depending on the context, Δ could be a function or, more generally, a transition probability. The transformation Δ generalizes the usual concept of pseudo-true value, as defined e.g. in Huber [23], Sawa [31], White [33] or Gouriéroux, Monfort and Trognon [18], and provides a reinterpretation of an estimated parameter $\mathbf{b} \in B$ (implicitly) associated with N in terms of an estimated parameter $\mathbf{a} \in A$ associated with M.

Limiting arguments lead to a concept of asymptotic encompassing. At this level, we retrieve the heuristic notion that 'valid' models on S, including the DGP itself, encompass all estimation procedures on S.

In general, exact encompassing will not hold, so we introduce a concept of specificity, dual to that of deficiency among sampling processes: see Lecam [26] and Cziszar [7]. Various measures of the specificity of N relative to $\Delta \circ M$ are considered. Of special interest are measures associated with conventional IPs, such as maximum likelihood (ML) estimation or Bayesian inference. Insofar as such measures depend on the sample $s \in S$, they are interpreted as measures of 'conditional' specificity (and are instrumental in the construction of a variety of encompassing test statistics). Unconditional measures of the specificity of N relative to $\Delta \circ M$ are defined as expectations of conditional specificity measures and require the introduction of a probability measure P_S^0 on $(S, \mathcal{S})$. Though the concept of (unconditional) specificity is generic, specific choices for P_S^0 depend upon the underlying mode of analysis (classical versus Bayesian, parametric versus non-parametric,...).

Naturally, we have to discuss the selection of a pseudo-true value for the purpose of measuring specificity when exact encompassing does not hold. Different viewpoints will be considered. The heuristic notion of using a transition Δ which minimizes the specificity of N relative to $\Delta \circ M$ typically leads to intractable functional optimization problems. The use of an asymptotic pseudo-true value often leads to major simplifications. Other choices based on asymptotic properties are also available.

One semantic issue requires clarification. As already discussed, encompassing and specificity fundamentally relate to 'results' (i.e. inference procedures) rather than to the underlying models themselves. This is true, in particular, for a model $\mathcal{N}$ to be encompassed, since $\mathcal{N}$ is inherently 'mis-specified' from the viewpoint of the encompassing model $\mathcal{M}$. In other words, $\mathcal{N}$ is essentially instrumental in the selection of an IP N whose outcome has to be accounted for within the context of an IP M associated with $\mathcal{M}$. As far as $\mathcal{M}$ is concerned, however, concepts such as unconditional specificity necessitate the introduction of a probability measure on $(S, \mathcal{S})$ whose choice is paired with that of $\mathcal{M}$ itself. To avoid constant reference to that distinction and to facilitate comparisons with earlier contributions, we discuss encompassing and specificity in terms of 'inferential models', i.e. in terms of pairs consisting of a sampling probability and an inference procedure, notwithstanding the fact that the former might serve no other purpose than that of rationalizing the selection of the latter.

The paper is organized as follows: section 2 provides an heuristic introduction to the concepts of encompassing and specificity, first from a classical viewpoint and then from a Bayesian perspective; technical concepts such as transition probabilities, inferential models and sufficiency are introduced in section 3; the concept of 'exact' encompassing is analyzed in section 4; lack of encompassing or specificity is discussed in section 5, together with related issues such as encompassing tests and a comparison between encompassing and model choice; section 6 considers asymptotic encompassing; approximate solutions to the frequently intractable concept of specificity are offered in section 7; the various concepts discussed in the paper are applied to the 'choice of regressors' problem in section 8 and section 9 concludes.

2 Encompassing and Specificity: an heuristic approach

To provide intuition for the formal definitions offered in the rest of the paper, we discuss encompassing and specificity at an heuristic level, first from a sampling theory viewpoint and then from a Bayesian perspective. Technical conditions - such as regularity conditions - are omitted for ease of discussion.

2.1 Classical Estimation

The relevant notation is collected in table 1.

Table 1: Classical notation

Model	$\mathcal{M}$		$\mathcal{N}$
parameter	$a \in A$		$b \in B$
sample		$s \in S$	
sampling density	$p(s\|a)$		$q(s\|b)$
estimators	$\tilde{a}(s)$		$\tilde{b}(s)$
pseudo-true value		$\beta(a)$	
estimated model	$\tilde{\mathcal{M}} = (\mathcal{M}, \tilde{a})$		$\tilde{\mathcal{N}} = (\mathcal{N}, \tilde{b})$

We first discuss finite sample situations and say that $\tilde{\mathcal{M}}$ exactly encompasses $\tilde{\mathcal{N}}$ if there exists a function $\beta : A \to B$ such that:

$$\tilde{b}(s) = \beta(\tilde{a}(s)) \quad s\text{-almost surely} \tag{2.1}$$

relative to $p(s|a)$, in which case $\tilde{b}$ can be obtained directly from $\tilde{a}$ without further processing of s. Condition (2.1) is strong and is only expected to hold under special circumstances.

Example 2.1: Let $S = \left\{\mathbf{y}_i \in \mathbb{R}^2; i = 1,\ldots,n\right\}$ consisting of n iid draws from $\mathsf{N}_2(\mathbf{a}, \Sigma)$ with $\Sigma = (\sigma_{ij})$ known. The ML estimator of $\mathbf{a}$ is given by the sample mean, $\tilde{\mathbf{a}}(s) = \bar{\mathbf{y}}$. Under $\mathcal{N}$, the mean vector $\mathbf{a}'$ is replaced by $(b : 0)$. The ML estimator of b in $\mathcal{N}$ is given by $\tilde{b}(s) = \boldsymbol{\pi}'\bar{\mathbf{y}}$ when $\boldsymbol{\pi}' = (1 : \sigma_{12}/\sigma_{22})$ is known. Then $\tilde{\mathcal{M}}$ exactly encompasses $\tilde{\mathcal{N}}$ with $\beta(\mathbf{a}) = \boldsymbol{\pi}'\mathbf{a}$ ■

Exact encompassing in the sense of (2.1) has two key characteristics:

(i) it is a transitive concept;

(ii) it is a relationship among estimators or, if models and estimators are paired through estimation principles (such as ML in example 2.1), among estimated models, not among the models themselves.

That (2.1) holds for example 2.1 is obviously related to the fact that $\mathcal{N}$ is 'nested' within $\mathcal{M}$ but only because ML estimators preserve nesting. There exist estimators, such as sample medians, for which (2.1) does not hold even though $\mathcal{N}$ is nested within $\mathcal{M}$. It is in order to avoid such confusion that encompassing and specificity are discussed in terms of estimated models, notwithstanding the deeper motivation that the main usage of a concept such as encompassing has always been one of accounting for 'results' or 'findings'. Thus, although exact encompassing is related to parsimonious encompassing (see [22]), the two concepts do not coincide.

If, as expected in most cases, (2.1) does not hold, we consider measuring a 'divergence' between $\tilde{b}(s)$ and $\beta(\tilde{a}(s))$ for a given pseudo-true value β, whose choice is discussed below. For example, if A and B are finite dimensional Euclidean spaces, we can use a norm such as:

$$d_H(s) = \left[\tilde{b}(s) - \beta(\tilde{a}(s))\right]' \mathbf{H} \left[\tilde{b}(s) - \beta(\tilde{a}(s))\right] \tag{2.2}$$

where $\mathbf{H}$ is a matrix function of s. This expression will be interpreted as a measure of the 'conditional specificity' of $\tilde{\mathcal{N}}$ relative to $\tilde{\mathcal{M}}$ with respect to β and can be used as a statistic for testing the hypothesis that $\tilde{\mathcal{M}}$ (asymptotically) encompasses $\tilde{\mathcal{N}}$ or, in light of the discussion which follows, for testing the 'validity' of $\tilde{\mathcal{M}}$ (in the direction of $\tilde{\mathcal{N}}$).

A measure of the 'unconditional specificity' of $\tilde{\mathcal{N}}$ relative to $\tilde{\mathcal{M}}$ is obtained by taking the expectation of d_H in (2.2) with respect to s under $\mathcal{M}$. It depends on β and we naturally consider selecting a β which minimizes the specificity of $\tilde{\mathcal{N}}$ relative to $\tilde{\mathcal{M}}$, though other criteria based, in particular, on asymptotic considerations may lead to more operational expressions.

Within the context of (pseudo) ML estimation, an alternative measure of the conditional specificity of $\tilde{\mathcal{N}}$ relative to $\tilde{\mathcal{M}}$ is:

$$d_L(s) = \int \log\left[\frac{q\left(t|\tilde{b}(s)\right)}{q\left(t|\beta\left(\tilde{a}(s)\right)\right)}\right] p(t|a)dt \tag{2.3}$$

Minimizing d_L with respect to β for all s is equivalent to solving the minimization problem:

$$\beta(a) = \underset{b}{\operatorname{argmin}} \int \log \left[\frac{p(s|a)}{q(s|b)}\right] p(s|a)ds \tag{2.4}$$

whose solution is given by the classical pseudo-true value associated with pseudo ML estimation.

We now briefly discuss the asymptotic case. Index the estimators by the sample size n and let n tend to infinity. Consider first the limiting form of (2.1) on $\mathcal{M}$. Let $a = \operatorname{plim}_{n\to\infty} \tilde{a}_n(s)$ and $\beta(a) = \operatorname{plim}_{n\to\infty} \tilde{b}_n(s)$ on $\mathcal{M}$. Then (2.1) holds asymptotically on $\mathcal{M}$ with β being a classical pseudo-true value. However, in contrast to exact encompassing, asymptotic encompassing is not transitive. The contradiction is only apparent and arises from the fact that while the finite sample distribution of $\tilde{a}_n(s)$ on $\mathcal{M}$ and $\tilde{b}_n(s)$ on $\mathcal{N}$ are typically 'equivalent' (i.e. have common null sets) which suffices to ensure the transitivity of (2.1), their limiting distributions are mutually 'singular'.

Insofar as $\mathcal{M}$ (as well as $\mathcal{N}$) is expected to be mis-specified relative to the DGP $\mathcal{P}$, we can usefully examine the limit of (2.1) on $\mathcal{P}$ rather than on $\mathcal{M}$. Let $a(\theta) = \operatorname{plim} \tilde{a}_n(s)$ and $b(\theta) = \operatorname{plim} \tilde{b}_n(s)$ on $\mathcal{P}$. Asymptotic encompassing now requires the existence of a function β such that $b(\theta) = \beta(a(\theta))$ for all θs, and so is not expected to hold in general. However, if (i) $A \subset \Theta$; (ii) $a(\theta) \equiv a$ and (iii) $b(\theta) \equiv \beta(a)$, then $\tilde{\mathcal{M}}$ asymptotically encompasses $\tilde{\mathcal{N}}$ relative to the usual pseudo-true value. We retrieve the heuristic notion that the (estimated) DGP (or any 'valid' reduction of it which is 'sufficient' relative to $\tilde{\mathcal{N}}$) asymptotically encompasses all rival models. This property sustains the use of encompassing tests for the 'validity' of $\tilde{\mathcal{M}}$ in the direction of $\tilde{\mathcal{N}}$.

2.2 Bayesian inference

The notation in table 2 complements that in table 1.

Table 2: Bayesian notation

Model	$\mathcal{M}$		$\mathcal{N}$
Prior density	$\mu(a)$		$\nu(b)$
Joint density	$\pi(s,a)$		$\chi(s,b)$
Predictive density	$p(s)$		$q(s)$
Posterior density	$\mu(a\|s)$		$\nu(b\|s)$
Transition density		$\delta(b\|a)$	
Inferential model	$\tilde{\mathcal{M}} = (\mathcal{M}, \mu(a\|s))$		$\tilde{\mathcal{N}} = (\mathcal{N}, \nu(b\|s))$

The Bayesian extension of (2.1) is straightforward. We say that $\tilde{\mathcal{M}}$ exactly encompasses $\tilde{\mathcal{N}}$ if there exists a conditional (transition) density $\delta(b|a)$, independent of s, such that:[1]

$$\nu(b|s) = \int_A \mu(a|s)\delta(b|a)da \quad s-a.s. \tag{2.5}$$

Example 2.1 (continued): Let the relevant prior densities be $\mathbf{a} \sim \mathsf{N}_2\left(\mathbf{a}_0, \mathbf{H}_0^{-1}\right)$ and $b \sim \mathsf{N}_1\left(b_0, l_0^{-1}\right)$. Let $\mathbf{H} = \Sigma^{-1} = (h_{ij})$. The corresponding posterior densities are $\mathbf{a}|s \sim \mathsf{N}_2\left(\mathbf{a}_*, \mathbf{H}_*^{-1}\right)$ and $b|s \sim \mathsf{N}_1\left(b_*, l_*^{-1}\right)$, where:

$$\mathbf{H}_* = \mathbf{H}_0 + n\mathbf{H} \qquad \mathbf{a}_* = \mathbf{H}_*^{-1}\left[n\mathbf{H}\bar{\mathbf{y}} + \mathbf{H}_0\mathbf{a}_0\right]$$
$$l_* = l_0 + nh_{11} \qquad b_* = l_*^{-1}\left[n(h_{11} : h_{12})\bar{\mathbf{y}} + l_0 b_0\right]$$

and $\bar{\mathbf{y}}' = (\bar{y}_1 : \bar{y}_2)$. Finally, let $b|\mathbf{a} \sim \mathsf{N}_1(\boldsymbol{\pi}'\mathbf{a}, v^2)$. Condition (2.5) holds if $\boldsymbol{\pi}'\mathbf{a}_* = b_*$ and $v^2 + \boldsymbol{\pi}'\mathbf{H}_*^{-1}\boldsymbol{\pi} = l_*^{-1}$, s-almost surely, i.e. if:

$$\text{(i) } \boldsymbol{\pi}' = l_*^{-1}(h_{11}^* : h_{12}^*)$$
$$\text{(ii) } l_0 b_0 = (h_{11}^0 : h_{12}^0)\mathbf{a}_0$$
$$\text{(iii) } v^2 = l_*^{-2}(l_0 - h_{11}^0)$$

which requires in particular that $l_0 \geq h_{11}^0$. If $\mu(\mathbf{a})$ and $\nu(b)$ are mutually 'consistent' with the nesting of $\mathcal{N}$ within $\mathcal{M}$, i.e. if $\nu(b)$ coincides with $\mu(a_1|a_2 = 0)$, then $l_0 = h_{11}^0$ and $b_0 = (1 : h_{11}^{0-1}h_{12}^0)\mathbf{a}_0$ so that conditions (i)-(iii) are verified with $v^2 = 0$ and $\boldsymbol{\pi}' = (1 : h_{11}^{*^{-1}}h_{12}^*)$. If $\mu(\mathbf{a})$ and $\nu(b)$ are 'non-informative' in the sense that $\mathbf{H}_0 = \mathbf{0}$ and $l_0 = 0$, then conditions (i)-(iii) still hold with $v^2 = 0$ and $\boldsymbol{\pi}' = (1 : h_{11}^{-1}h_{12})$ in which case $\delta(b|\mathbf{a})$ collapses to a Dirac transition probability on the classical pseudo-true value $\beta(\mathbf{a})$. ■

The comments made earlier extend to the Bayesian case. Classical pseudo-true values which are functions from A to B are now replaced by Bayesian pseudo-true values which are transition probabilities from A to $(B, \mathcal{B})$. The Bayesian concept of encompassing calls for a number of additional comments.

First, (2.5) involves two parameters (a, b) and one statistic s. Consider instead two statistics (s, t) and one parameter a, and substitute (s, t, a) for (a, b, s) in (2.5), adjusting notation to eliminate ambiguities. This substitution yields the following formula:

$$q(t|a) = \int_S p(s|a)\lambda(t|s)ds \tag{2.6}$$

where q denotes the sampling density of t and λ is a conditional density for t, given s, *independent of* a. Then (2.6) corresponds to a version, expressed in terms

[1] Requiring (2.5) to hold s-almost surely is tantamount to requiring that s be independent of b, conditionally on a. As discussed below, that condition makes sense from the viewpoint of $\mathcal{M}$. It does not contradict the 'likelihood principle', whereby all inferences should be conditional on the actual sample s_*. As discussed in section [5] Bayesian tests of whether or not (2.5) holds are evaluated at s_*. Moreover, requiring (2.5) to hold only at s_* would empty the concept of encompassing of meaning since it would be trivially satisfied by the transition $\delta(b|a, s_*) = \nu(b|s_*)$.

of density functions, of the definition of sufficiency (among statistics) defined by Blackwell [5], [6]. The 'duality' between (2.5) and (2.6) sustains our interpretation of encompassing as a notion of sufficiency among models.

Secondly, the conditional density $\delta(b|a)$ *de facto* generates an extension of the joint density π on $S \times A$ associated with $\mathcal{M}$, into a density π^* on $S \times A \times B$ such that π is a marginal of π^*, thereby preserving all the features of $\mathcal{M}$. The density π^* is defined as:

$$\begin{aligned} \pi^*(s,a,b) &= [p(s|a)\,\mu(a)]\,\delta(b|a) \\ &= [\mu(a|s)p(s)]\,\delta(b|a) \end{aligned} \tag{2.7}$$

Let the superscript * denote marginal and conditional densities associated with π^*. In particular, the sampling distribution associated with π^* is:

$$p^*(s|a,b) = p(s|a) \tag{2.8}$$

so that π^* incorporates the assumption that a is a sufficient parameterization (i.e. that s and b are independent conditionally on a), an assumption which is largely implicit in the formulation of $\mathcal{M}$ by its proprietor. Under π^*, the posterior distribution of b is given by:

$$\nu^*(b|s) = \int_A \mu(a|b)\delta(b|a)da \tag{2.9}$$

Hence, (2.5) essentially requires that the 'actual' posterior density $\nu(b|s)$, as initially obtained within $\mathcal{N}$, coincides with the 'derived' posterior density $\nu^*(b|s)$, which is obtained within $\mathcal{M}$ via the transition $\delta(\cdot)$.

Extensions of the concept of specificity to the Bayesian case are fairly straightforward and are discussed below within a general framework. Again the issue arises of which transition $\delta(\cdot)$ ought to be used for the purpose of measuring specificity. A strict decisional approach would require that the proprietor of $\mathcal{M}$ be capable of eliciting a genuine *joint* prior on a and b, wherefrom $\delta(\cdot)$ would follow by conditioning. Such an exercise is demanding and requires a thorough understanding of the (stochastic) relationship between a and b.[2] Further, it generates a measure of specificity which is problem dependent. Our objective is to evaluate $\tilde{\mathcal{M}}$ relative to $\tilde{\mathcal{N}}$ without recourse to such complex elicitation exercises. Hence we propose instead to select a transition $\delta(\cdot)$ which minimizes the predictive expectation of an appropriate measure of divergence between $\nu(b|s)$ and $\nu^*(b|s)$. Specificity is then defined as a lower bound to the expected divergence between $\nu(b|s)$ and $\nu^*(b|s)$

[2] A similar problem arises in the Bayesian literature on model choice and is often 'addressed' by assuming prior independence between a and b. This default option is unsatisfactory as it is incompatible with the concept of encompassing. The relationship between measures of specificity and posterior odds is formally investigated in section 5.5.

and is meant to measure some 'irreducible divergence' between $\tilde{\mathcal{M}}$ and $\tilde{\mathcal{N}}$.[3] This notion of a minimal expected divergence is precisely that which has already been used at a dual level for the purpose of measuring lack of sufficiency, i.e. deficiency in the terminology of Lecam [26] and Cziszar [7].

Finally, measures of divergence between $\nu(b|s)$ and $\nu^*(b|s)$ are functions of the actual sample s. Hence any such measure can be used for the purpose of constructing a Bayesian test of whether or not $\tilde{\mathcal{M}}$ encompasses $\tilde{\mathcal{N}}$, following the testing principles discussed in Florens and Mouchart [12]. Specifically, we can use the predictive densities associated with $\mathcal{M}$ and $\mathcal{N}$ respectively as null and alternative hypotheses to evaluate the 'significance' of the actual encompassing test statistics. Additional details are provided in section 5.4.

2.3 Comments

It should now be clear that the classical and Bayesian concepts of encompassing have much in common. In fact, there exists a technical concept which can reconcile both viewpoints, namely that of a transition probability. The adoption of that concept, which entails rephrasing the analysis in terms of a more abstract probability framework, generates additional advantages: by focusing attention on the deeper probability structure of the competing models, it eliminates several side issues that may create confusion in a less general framework (non-uniqueness of the parameterization associated with any given model, redundant parameters, singular distributions, and so on).

3 Preliminaries

To be self-contained, we next describe the technical concepts used in the rest of the paper. Details can be found e.g. in Neveu [29], Dellacherie and Meyer [9], and Florens *et al.* [13] whose (Bayesian) framework is ideally suited to the object of our paper. The reader may wish to skim through section 3.1 since a thorough understanding of notation is only required for proofs. Understanding our definitions and the main results essentially requires familiarity with formula (3.9) below or, for heuristics, with its density counterpart, as given in (3.12).

[3] In line with the recent literature on Bayesian robustness, as discussed e.g in Berger [1] or Lavine [25], we could alternatively consider computing lower and upper bounds to measures of divergence between $\nu(b|s)$ and $\nu^*(b|s)$ within a given class of transition probabilities. This suggestion will not be investigated further in the present paper but belongs to our research agenda.

3.1 Transition probabilities

The concept of transition probability (equivalently Markov kernel or random function) is central to the argument. Hence, we summarize its main properties using the notation in Neveu [29](section III.2) to which the reader is referred for more details. A short-hand notation will be introduced at the end of this section. Let $(U,\mathcal{U})$, and $(V,\mathcal{V})$ denote two measurable spaces. Let $[\mathcal{U}]_\infty$ and $[\mathcal{V}]_\infty$ denote the corresponding sets of bounded random variables.[4]

Definition 3.1. *A transition probability is a function:*[5]

$$\Lambda^{\mathcal{U}}_{\mathcal{V}} : U \times \mathcal{V} \rightarrow [0,1];\ (u,Y) \rightarrow \Lambda^{\mathcal{U}}_{\mathcal{V}}(u,Y) \tag{3.1}$$

which has the following properties:
(i) $\forall u \in U,\ \Lambda^{\mathcal{U}}_{\mathcal{V}}(u,\cdot)$ *is a probability on* $(V,\mathcal{V})$;
(ii) $\forall Y \in \mathcal{V},\ \Lambda^{\mathcal{U}}_{\mathcal{V}}(\cdot,Y)$ *is* $\mathcal{U}$*-measurable.*

We make use of the following properties of transition probabilities:
(i) To every pair consisting of a probability $P_{\mathcal{U}}$ on $(U,\mathcal{U})$ and of a transition probability $\Lambda^{\mathcal{U}}_{\mathcal{V}}$ on $U \times \mathcal{V}$, we can associate a probability $\Pi_{\mathcal{U}\otimes\mathcal{V}}$ on the product space $(U \times V, \mathcal{U} \otimes \mathcal{V})$ and a probability $Q_{\mathcal{V}}$ on $(V,\mathcal{V})$, respectively defined by:

$$\forall X \in \mathcal{U}, Y \in \mathcal{V}, \quad \Pi_{\mathcal{U}\otimes\mathcal{V}}(X \times Y) = \int_X \Lambda^{\mathcal{U}}_{\mathcal{V}}(u,Y)P_{\mathcal{U}}(du) \tag{3.2}$$

$$\forall Y \in \mathcal{V}, \quad Q_{\mathcal{V}}(Y) = \int_U \Lambda^{\mathcal{U}}_{\mathcal{V}}(u,Y)P_{\mathcal{U}}(du) \tag{3.3}$$

(ii) To every pair consisting of a random variable $y \in [\mathcal{V}]_\infty$ and of a transition probability $\Lambda^{\mathcal{U}}_{\mathcal{V}}$ on $U \times \mathcal{V}$, we can associate a random variable $x \in [\mathcal{V}]_\infty$ defined as:

$$\forall u \in \mathcal{U}, \quad x(u) = \int_V y(v)\Lambda^{\mathcal{U}}_{\mathcal{V}}(u,dv) \tag{3.4}$$

(iii) To every pair consisting of a transition probability $\Lambda^{\mathcal{U}}_{\mathcal{V}}$ on $U \times \mathcal{V}$ and a transition probability $\Delta^{\mathcal{V}}_{\mathcal{W}}$ on $V \times \mathcal{W}$, we can associate a transition probability $\Gamma^{\mathcal{U}}_{\mathcal{W}}$ on $U \times \mathcal{W}$ defined as:

$$\forall Z \in \mathcal{W}, \quad \Gamma^{\mathcal{U}}_{\mathcal{W}}(u,Z) = \int_V \Delta^{\mathcal{V}}_{\mathcal{W}}(v,Z)\Lambda^{\mathcal{U}}_{\mathcal{V}}(u,dv) \tag{3.5}$$

In the rest of the paper, we use a short-hand notation taken from Florens *et al.* [13](Ch.0) which leads to the following reformulation of formulae (3.3)-(3.5):

$$\forall Y \in \mathcal{V}, \quad Q_{\mathcal{V}}(Y) = \int_U \Lambda^{\mathcal{U}}_{\mathcal{V}}(Y)dP_{\mathcal{U}} \tag{3.6}$$

[4] The restriction to bounded random variables is introduced for convenience, since such variables are integrable under any probability measures. In practice, we will consider much larger classes of random variables depending on the specific probability measures which are being used.

[5] The notation $\Lambda^{\mathcal{U}}_{\mathcal{V}}$ is adopted to emphasize the measurability requirement.

$$x = \int_V y d\Lambda_{\mathcal{V}}^{\mathcal{U}} \tag{3.7}$$

$$\forall Z \in \mathcal{W}, \quad \Gamma_{\mathcal{W}}^{\mathcal{U}}(Z) = \int_V \Delta_{\mathcal{W}}^{\mathcal{V}}(Z) d\Lambda_{\mathcal{V}}^{\mathcal{U}} \tag{3.8}$$

Hence, the notation $\Lambda_{\mathcal{V}}^{\mathcal{U}}$ covers several usages: it represents either the transition probability itself, or a mapping from a set of probabilities on $(U,\mathcal{U})$ onto a set of probabilities on $(V,\mathcal{V})$ as in (3.6), or a mapping from $[\mathcal{V}]_\infty$ onto $[\mathcal{U}]_\infty$ as in (3.7). However, no ambiguity should arise from this multiplicity of usages since, in particular, formulae such as (3.6)-(3.8) are unequivocal.

The third interpretation, whereby (3.7) asserts that $x \in [\mathcal{U}]_\infty$ is the image of $y \in [\mathcal{V}]_\infty$ by the mapping $\Lambda_{\mathcal{V}}^{\mathcal{U}}$, offers the advantage that (3.8) then corresponds to the usual composition for mappings. Specifically:

$$\text{If} \quad x = \Lambda_{\mathcal{V}}^{\mathcal{U}}(y) \quad \text{and} \quad y = \Delta_{\mathcal{W}}^{\mathcal{V}}(z), \quad \text{then} \quad x = \Gamma_{\mathcal{W}}^{\mathcal{U}}(z)$$

with

$$\Gamma_{\mathcal{W}}^{\mathcal{U}} = \Lambda_{\mathcal{V}}^{\mathcal{U}} \circ \Delta_{\mathcal{W}}^{\mathcal{V}} \tag{3.9}$$

A proof that (3.8) and (3.9) are equivalent relies upon monotone class arguments and is found e.g. in Dellacherie and Meyer [9]. The more compact formulation (3.9) is used in the rest of the paper.

Under suitable dominance arguments,[6] we can associate bimeasurable density functions with transition probabilities and, for example, rewrite (3.6)-(3.8) in terms of densities as:

$$q(v) = \int_U \lambda(v|u)p(u)du \tag{3.10}$$

$$x(u) = \int_V y(v)\lambda(v|u)dv \tag{3.11}$$

$$\gamma(w|u) = \int_V \delta(w|v)\lambda(v|u)du \tag{3.12}$$

Such reformulations are useful for heuristic arguments but not for formal proofs.

One class of transition probabilities which plays an important role in the analysis of the limiting behavior of posterior distributions is the class of Dirac transition

[6] A transition probability $\Lambda_{\mathcal{V}}^{\mathcal{U}}$ is said to be dominated if there exists a σ-finite measure on $(V,\mathcal{V})$ such that for all $u \in U, \Lambda_{\mathcal{V}}^{\mathcal{U}}(u,\cdot)$ is dominated by that measure. Under suitable regularity conditions (see Florens *et al.* [13], theorem 0.3.19), there will exist a bimeasurable function $\lambda(u,v)$ such that:

$$\Lambda_{\mathcal{V}}^{\mathcal{U}}(u,Y) = \int_Y \lambda(u,v)dv$$

where the integration is relative to the dominant measure. For notational convenience, we shall not introduce an additional symbol to denote the dominant measure.

probabilities. Specifically let $\lambda : (U,\mathcal{U}) \rightarrow (V,\mathcal{V})$ be a $\mathcal{U}$-measurable function. The corresponding Dirac transition probability is defined as:

$$\forall u \in U, \quad \forall Y \in \mathcal{V}, \quad D^{\mathcal{U}}_{\mathcal{V},\lambda}(u,Y) = \begin{cases} 0 & \text{if } \lambda(u) \notin Y \\ 1 & \text{if } \lambda(u) \in Y \end{cases} \tag{3.13}$$

3.2 Inferential models

To discuss sufficiency and encompassing in parallel, we need two sample spaces and two parameter spaces. The notation is shown in Table 3.

Table 3: Inference Notation

	Parameters		Samples	
Outcomes	$a \in A$	$b \in B$	$s \in S$	$t \in T$
Events	$G \in \mathcal{A}$	$F \in \mathcal{B}$	$X \in \mathcal{S}$	$Y \in \mathcal{T}$
Random Variables	$g \in [\mathcal{A}]_\infty$	$f \in [\mathcal{B}]_\infty$	$x \in [\mathcal{S}]_\infty$	$y \in [\mathcal{T}]_\infty$

A classical experiment is defined by a set of sampling probabilities indexed by a parameter. Bayesian reasoning endows the parameter space with a σ-field and hence implicitly reinterprets sampling probabilities as transition probabilities.

Definition 3.2. *A sampling model is a triple consisting of a measurable parameter space $(A, \mathcal{A})$, a measurable sample space $(S, \mathcal{S})$ and a transition (sampling) probability $P^{\mathcal{A}}_{S}$.*

Definition 3.3. *An inferential model $\mathcal{M}_M$ is a pair consisting of a sampling model $\mathcal{M}$ and an estimation procedure $M^{\mathcal{S}}_{\mathcal{A}}$.*

Definition 3.4. *A Bayesian inferential model $\mathcal{M}^{\mu}_{M}$ is a triple consisting of a sample model $\mathcal{M}$, a prior probability $\mu_{\mathcal{A}}$ and the corresponding posterior probability $\mu^{\mathcal{S}}_{\mathcal{A}}$.*

We consider two inferential models using the notation in table 4.

Table 4: Probability Notation

Sampling Model	$\mathcal{M} = \{(A,\mathcal{A}),(S,\mathcal{S}),P^{\mathcal{A}}_{S}\}$	$\mathcal{N} = \{(B,\mathcal{B}),(T,\mathcal{T}),Q^{\mathcal{B}}_{T}\}$
Prior Probabilities	$\mu_{\mathcal{A}}$	$\nu_{\mathcal{B}}$
Joint Probabilities	$\Pi_{\mathcal{A}\otimes\mathcal{S}}$ or Π	$\chi_{\mathcal{B}\otimes\mathcal{T}}$ or χ
Predictive Probabilities	$P_{\mathcal{S}}$	$Q_{\mathcal{T}}$
Posterior Probabilities	$\mu^{\mathcal{S}}_{\mathcal{A}}$	$\nu^{\mathcal{T}}_{\mathcal{B}}$
Estimation Procedures	$M^{\mathcal{S}}_{\mathcal{A}}$	$N^{\mathcal{T}}_{\mathcal{B}}$
Inferential Models	$\mathcal{M}_M = (\mathcal{M}, M^{\mathcal{S}}_{\mathcal{A}})$	$\mathcal{N}_N = (\mathcal{N}, N^{\mathcal{T}}_{\mathcal{B}})$
Bayes Inferential Models	$\mathcal{M}^{\mu}_{M} = (\mathcal{M}, \mu_{\mathcal{A}}, \mu^{\mathcal{S}}_{\mathcal{A}})$	$\mathcal{N}^{\nu}_{N} = (\mathcal{N}, \nu_{\mathcal{B}}, \nu^{\mathcal{T}}_{\mathcal{B}})$

This notation is used in the following 'dual context':[7]

(i) The concept of sufficiency applies to a pair of sampling models sharing a common parameter space ($\mathcal{A} = \mathcal{B}$) and, as shown in section 3.3, specifically relates the two sampling probabilities $P_{\mathcal{S}}^{\mathcal{A}}$ and $Q_{\mathcal{T}}^{\mathcal{A}}$.

(ii) Encompassing applies instead to a pair of models sharing a common sample space ($\mathcal{S} = \mathcal{T}$) and relates together two arbitrary estimation procedures, say $M_{\mathcal{A}}^{\mathcal{S}}$ and $N_{\mathcal{B}}^{\mathcal{S}}$. Examples of estimation procedures are:

(a) Estimators: if $\hat{a} : (S, \mathcal{S}) \rightarrow (A, \mathcal{A})$ is an 'estimator', then the Dirac measure $D_{\mathcal{A},\hat{a}}^{\mathcal{S}}$ is an estimation procedure ;

(b) Estimated sampling distributions: if an estimator $\hat{a}$ has a sampling distribution $\phi(a)$, then $\phi(\hat{a})$ defines an estimation procedure;

(c) Posterior distributions: if $\mathcal{M}$ is endowed with a prior density $\mu_{\mathcal{A}}$, then the corresponding posterior density $\mu_{\mathcal{A}}^{\mathcal{S}}$ is an estimation procedure.

3.3 Sufficiency

The sufficiency concept to which encompassing is related by duality was introduced by Blackwell [5], [6] and is extensively analyzed in Lecam [26]: also see Goel and DeGroot [16] and Torgensen [32]. The classical definition is:

Definition 3.5. *Let $\mathcal{M}$ and $\mathcal{N}$ be two sampling models with a common parameter space ($\mathcal{A} = \mathcal{B}$). $\mathcal{M}$ is sufficient for $\mathcal{N}$ if and only if there exists a transition probability $\Lambda_{\mathcal{T}}^{\mathcal{S}}$ such that:*

$$Q_{\mathcal{T}}^{\mathcal{A}} = P_{\mathcal{S}}^{\mathcal{A}} \circ \Lambda_{\mathcal{T}}^{\mathcal{S}} \tag{3.14}$$

If, in a Bayesian framework, a common prior probability $\mu_{\mathcal{A}}$ is associated with the two sampling models, then the sufficiency condition has to hold $\mu_{\mathcal{A}}$-almost surely. More generally, definition 3.5 can be reformulated in several ways under equivalent priors.[8] In particular, we can enlarge the sampling model $\mathcal{M}$ into a sampling model $\mathcal{M}_e$ whose sampling probability $P_{\mathcal{S}\otimes\mathcal{T}}^{\mathcal{A}}$ is an 'extension' of $P_{\mathcal{S}}^{\mathcal{A}}$ defined such that:

(i) $\mathcal{S}$ is sufficient or, equivalently, $\mathcal{T} \perp\!\!\!\perp \mathcal{A} | \mathcal{S}$ (i.e. $\mathcal{T}$ and $\mathcal{A}$ are independent, conditionally on $\mathcal{S}$) under $\mathcal{M}_e$;

(ii) $P_{\mathcal{S}\otimes\mathcal{T}}^{\mathcal{A}} | \mathcal{T}$ restricted to $\mathcal{T}$ equals $Q_{\mathcal{T}}^{\mathcal{A}}$.

See Florens *et al.* [13] for details and for discussion of the case where the two sampling models are endowed with non-equivalent prior probabilities.

[7] Here, predictive probabilities are marginal probabilities for the data (i.e. 'prior predictive' probabilities) rather than conditional probabilities for out-of-sample data given actual data (i.e. 'posterior predictive' probabilities). The omission of 'prior' as a qualifier should not cause any confusion.

[8] Two probabilities μ and μ' are equivalent if $\forall A \in \mathcal{A}, \mu(A) = 0 \leftrightarrow \mu'(A) = 0$.

4 Exact Encompassing

4.1 General Definitions

Our baseline definition of exact encompassing is the dual of definition 3.5 and applies to arbitrary estimation procedures. The two inferential models under consideration share a common sample space. Hence, the notation in table 4 applies with $\mathcal{S} = \mathcal{T}$.

As noted in the introduction, an important qualification applies to all definitions and results which follow, namely that they assume identities that are conditional on $\mathcal{S}$ (or on sub σ-fields thereof) and are meant to be almost sure with respect to a 'reference' probability $P_{\mathcal{S}}^0$ on $(S, \mathcal{S})$. The choice of $P_{\mathcal{S}}^0$, which essentially serves to characterize the relevant null sets, often depends on the context. Natural choices are $P_{\mathcal{S}}^{\mathcal{A}}$ from a classical viewpoint or $P_{\mathcal{S}}$ from a Bayesian perspective. In line with the recent econometric literature on 'mis-specified' models - see e.g. Gouriéroux *et al.* [17] - we could also think of $P_{\mathcal{S}}^0$ as representing the underlying DGP.

Definition 4.1. *Let $\mathcal{M}_M$ and $\mathcal{N}_N$ be two inferential models. $\mathcal{M}_M$ exactly encompasses $\mathcal{N}_N$ (on $P_{\mathcal{S}}^0$) if and only if there exists a transition probability $\Delta_{\mathcal{B}}^{\mathcal{A}}$, called the pseudo-true value of $\mathcal{N}_N$ within $\mathcal{M}_M$, such that:*

$$N_{\mathcal{B}}^{\mathcal{S}} = M_{\mathcal{A}}^{\mathcal{S}} \otimes \Delta_{\mathcal{B}}^{\mathcal{A}}, \qquad P_{\mathcal{S}}^0 \text{ - a.s.} \tag{4.1}$$

Lemma 4.1. *If $\mathcal{M}_M$ exactly encompasses $\mathcal{N}_N$ (on $P_{\mathcal{S}}^0$) with pseudo-true value $\Delta_{\mathcal{B}}^{\mathcal{A}}$, if $\mathcal{N}_N$ exactly encompasses $\mathcal{O}_O = (\mathcal{O}, O_{\mathcal{C}}^{\mathcal{S}})$ (on $Q_{\mathcal{S}}^0$) with pseudo-true value $\Lambda_{\mathcal{C}}^{\mathcal{B}}$ and if $P_{\mathcal{S}}^0$ and $Q_{\mathcal{S}}^0$ are equivalent, then $\mathcal{M}_M$ exactly encompasses $\mathcal{O}_O$ (on $P_{\mathcal{S}}^0$) with pseudo-true value:*

$$\Gamma_{\mathcal{C}}^{\mathcal{A}} = \Delta_{\mathcal{B}}^{\mathcal{A}} \circ \Lambda_{\mathcal{C}}^{\mathcal{B}}, \qquad P_{\mathcal{S}}^0 \text{ - a.s.} \tag{4.2}$$

■

Proof: Follows from (3.5).

Lemma 4.1 establishes that exact encompassing is transitive. If $\mathcal{N}_N$ is encompassed by $\mathcal{M}_M$, its status need not be reexamined if $\mathcal{M}_M$ is later replaced by an encompassing model $\mathcal{O}_O$.

Example 4.1: The concept of parametric encompassing, as defined e.g. in Mizon and Richard [28], applies to situations where the estimation procedures $M_{\mathcal{A}}^{\mathcal{S}}$ and $N_{\mathcal{B}}^{\mathcal{S}}$ are Dirac measures associated with a pair of estimators, $\tilde{a}$ and $\tilde{b}$ respectively. An *additional* restriction is imposed, namely that $\Delta_{\mathcal{B}}^{\mathcal{A}}$ is itself a Dirac measure.

Under these conditions, (4.1) simplifies to:

$$\begin{aligned}\forall f \in [\mathcal{B}]_\infty, \forall s \in S,\ \ f\left(\tilde{b}(s)\right) &= N_{\mathcal{B}}^{\mathcal{S}}(f) = \int\limits_A \left[\textstyle\int_B f d\Delta_{\mathcal{B}}^{\mathcal{A}}\right] dM_{\mathcal{A}}^{\mathcal{S}} \\ &= \int\limits_A f\left(\beta(a)\right) dM_{\mathcal{A}}^{\mathcal{S}} = \left(f \circ \beta\right)\left(\tilde{a}\left(s\right)\right)\end{aligned} \qquad (4.3)$$

which is essentially formula (2.4) in Mizon and Richard [28] (when $\tilde{\phi} = 0$). As discussed in section 2.1 adopting a limiting viewpoint, whereby $\beta(a)$ is a classical pseudo-true value, results in a loss of transitivity. ■

The next example is phrased in terms of sampling distributions but its Bayesian reformulation in terms of posterior densities is straightforward.

Example 4.2: Let $\tilde{\mathbf{a}}$ be an estimator which is $\mathsf{N}_k(\mathbf{a}, \mathbf{\Sigma}_a)$ under $\mathcal{M}$ and $\tilde{\mathbf{b}}$ an estimator which is $\mathsf{N}_\ell(\mathbf{b}, \mathbf{\Omega}_b)$ under $\mathcal{N}$. Let $\mathsf{N}_k(\tilde{\mathbf{a}}, \mathbf{\Sigma}_{\tilde{a}})$ and $\mathsf{N}_\ell\left(\tilde{\mathbf{b}}, \mathbf{\Omega}_{\tilde{b}}\right)$ be the corresponding estimation procedures. We restrict attention to linear Gaussian transition probabilities, so that

$$\Delta_{\mathcal{B}}^{\mathcal{A}} = \mathsf{N}_\ell(\mathbf{C}\mathbf{a} + \mathbf{c}, \mathbf{V}) \qquad (4.4)$$

Formula (4.1) then requires that there exist $\mathbf{C}, \mathbf{c}$ and a symmetric positive semi-definite matrix $\mathbf{V}$ such that $\forall s \in S$, $\tilde{\mathbf{b}} = \mathbf{C}\tilde{\mathbf{a}} + \mathbf{c}$ and $\mathbf{\Omega}_{\tilde{b}} = \mathbf{C}\mathbf{\Sigma}_{\tilde{a}}\mathbf{C}' + \mathbf{V}$. In such a case, $\mathbf{V}$ measures the loss of efficiency when $\mathbf{C}\mathbf{a}+\mathbf{c}$ is estimated by $\tilde{\mathbf{b}}$ in $\mathcal{N}$ instead of $\mathbf{C}\tilde{\mathbf{a}} + \mathbf{c}$ in $\mathcal{M}$. ■

The property of exact encompassing may be weakened in two non-mutually exclusive directions:

(i) we may consider only a sub σ-field of $\mathcal{B}$ consisting of events of special interest within the context of $\mathcal{N}$ (partial encompassing);

(ii) we may also condition the entire analysis on a sub σ-field $\mathcal{S}$ consisting e.g. of events relative to a set of 'exogenous' variables and, in particular, let the pseudo-true values be conditional on that sub σ-field.

Let $\mathcal{B}_1$ and $\mathcal{S}_1$ be sub σ-fields of $\mathcal{B}$ and $\mathcal{S}$ respectively.

Definition 4.2. *The inferential model $\mathcal{M}_M$ exactly encompasses the inferential model $\mathcal{N}_N$ on $\mathcal{B}_1$ given $\mathcal{S}_1$ (on P_S^0) if and only if there exists a transition probability $\Delta_{\mathcal{B}_1}^{\mathcal{A}\otimes\mathcal{S}_1}$ such that:*

$$N_{\mathcal{B}_1}^{\mathcal{S}} = M_{\mathcal{A}}^{\mathcal{S}} \circ \Delta_{\mathcal{B}_1}^{\mathcal{A}\otimes\mathcal{S}_1} \qquad P_S^0\text{-a.s.} \qquad (4.5)$$

Example 4.3: The conventional 'choice of regressors' problem typically takes the form:

$$\mathcal{M}\ :\quad \mathrm{y} = \mathbf{X}\boldsymbol{\beta} + \mathrm{u}, \quad \mathrm{u} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_T\right), \quad \boldsymbol{\beta} \in \mathbb{R}^k, \quad \mathbf{a} = (\boldsymbol{\beta}, \sigma^2) \qquad (4.6)$$

$$\mathcal{N}\ :\quad \mathrm{y} = \mathbf{Z}\gamma + \mathrm{v}, \quad \mathrm{v} \sim \mathsf{N}\left(\mathbf{0}, \tau^2\mathbf{I}_T\right), \quad \gamma \in \mathbb{R}^\ell, \quad \mathrm{b} = (\gamma, \tau^2) \qquad (4.7)$$

where $\mathbf{X}$ and $\mathbf{Z}$ are conditioning variables. Let $\mathcal{B}_1$ and $\mathcal{S}_1$ be the sub σ-fields associated with γ and $(\mathbf{X},\mathbf{Z})$ respectively. In their discussion of parametric encompassing, Mizon and Richard [28] use the Dirac transition measure associated with the classical pseudo-true value, namely $\gamma_\beta = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\beta$. Bayesian generalizations are discussed in section 5.3. ■

Lemma 4.2. *If $\mathcal{M}_M$ exactly encompasses $\mathcal{N}_N$ on $\mathcal{B}_1$ given $\mathcal{S}_1$ (on $P_{\mathcal{S}}^0$) with pseudo-true value $\Delta_{\mathcal{B}_1}^{\mathcal{A}\otimes\mathcal{S}_1}$, then $\mathcal{M}_M$ exactly encompasses $\mathcal{N}_N$ on any $\mathcal{B}_0$ given any $\mathcal{S}_0$ such that $\mathcal{B}_0 \subset \mathcal{B}_1$ and $\mathcal{S}_1 \subset \mathcal{S}_0$ with pseudo-true value:*

$$\forall F \in \mathcal{B}_0,\ \Delta_{\mathcal{B}_0}^{\mathcal{A}\otimes\mathcal{S}_0}(F) = \int_F d\Delta_{\mathcal{B}_1}^{\mathcal{A}\otimes\mathcal{S}_1} \qquad P_{\mathcal{S}}^0\text{-a.s.} \tag{4.8}$$

Proof: F necessarily belongs to $\mathcal{B}_1$ and $\Delta_{\mathcal{B}_1}^{\mathcal{A}\otimes\mathcal{S}_1} = \Delta_{\mathcal{B}_1}^{\mathcal{A}\otimes\mathcal{S}_0}$ for any $\mathcal{S}_0$ such that $\mathcal{S}_1 \subset \mathcal{S}_0$. ■

It is often the case that $\mathcal{A}$ is a product space with $\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2$ and that the inference procedure $M_{\mathcal{A}}^{\mathcal{S}}$ accordingly factorizes such that:

$$M_{\mathcal{A}_1}^{\mathcal{S}} = M_{\mathcal{A}_1}^{\mathcal{S}_1} \quad \text{and} \quad M_{\mathcal{A}_2}^{\mathcal{S}\otimes\mathcal{A}_1} = M_{\mathcal{A}_2}^{\mathcal{S}} \tag{4.9}$$

Definition 4.3. *$\mathcal{M}_M^r = (\mathcal{M}, M_{\mathcal{A}_2}^{\mathcal{S}})$ is a valid reduction of $\mathcal{M}_M$ on $\mathcal{S}_1$ if condition (4.9) is satisfied.*

Lemma 4.3. *If $\mathcal{M}_M$ exactly encompasses $\mathcal{N}_N$ (on $P_{\mathcal{S}}^0$) with pseudo-true value $\Delta_{\mathcal{B}}^{\mathcal{A}}$, and if $\mathcal{M}_M^r$ is a valid reduction of $\mathcal{M}_M$ on $\mathcal{S}_1$, then $\mathcal{M}_M^r$ exactly encompasses $\mathcal{N}_N$ given $\mathcal{S}_1$ with pseudo-true value $\Delta_{\mathcal{B}}^{\mathcal{A}_2\otimes\mathcal{S}_1}$ given by:*

$$\Delta_{\mathcal{B}}^{\mathcal{A}_2\otimes\mathcal{S}_1} = M_{\mathcal{A}_1}^{\mathcal{S}_1} \circ \Delta_{\mathcal{B}}^{\mathcal{A}} \qquad P_{\mathcal{S}}^0\text{ -a.s.} \tag{4.10}$$

Proof: Under (4.1) and (4.9) we have successively:

$$\begin{aligned}\forall F \in \mathcal{B},\quad N_{\mathcal{B}}^{\mathcal{S}}(F) &= \int \Delta_{\mathcal{B}}^{\mathcal{A}}(F)\,dM_{\mathcal{A}}^{\mathcal{S}} \\ &= \int \Delta_{\mathcal{B}}^{\mathcal{A}_1\otimes\mathcal{A}_2}(F)\,dM_{\mathcal{A}_2}^{\mathcal{S}\otimes\mathcal{A}_1}\,dM_{\mathcal{A}_1}^{\mathcal{S}} \\ &= \int \Delta_{\mathcal{B}}^{\mathcal{A}_2\otimes\mathcal{S}_1}(F)\,dM_{\mathcal{A}_2}^{\mathcal{S}} \qquad P_{\mathcal{S}}^0\text{ -a.s.}\end{aligned}$$

■

Condition (4.9) is often associated with a factorization of the sampling probability $P_{\mathcal{S}}^{\mathcal{A}}$ into a marginal probability $P_{\mathcal{S}_1}^{\mathcal{A}}$ and a conditional one $P_{\mathcal{S}}^{\mathcal{A}\otimes\mathcal{S}_1}$ in such a way that:

$$P_{\mathcal{S}_1}^{\mathcal{A}} = P_{\mathcal{S}_1}^{\mathcal{A}_1} \quad \text{and} \quad P_{\mathcal{S}}^{\mathcal{A}\otimes\mathcal{S}_1} = P_{\mathcal{S}}^{\mathcal{A}_2\otimes\mathcal{S}_1} \tag{4.11}$$

If, for example, $\mathcal{M}_{\mathcal{A}}^{\mathcal{S}}$ is the posterior probability associated with the prior $\mu_{\mathcal{A}}$, then (4.11) together with the prior independence condition $\mathcal{A}_1 \perp\!\!\!\perp \mathcal{A}_2$ defines a global cut in the terminology adopted by Florens *et al.* [13] and (4.9) follows.

4.2 Bayesian exact encompassing

We now restrict attention to Bayesian inferential models. By taking advantage of the relationship between the prior and posterior probabilities, we can derive additional implications of exact encompassing. The two Bayesian inferential models under consideration are denoted $\mathcal{M}_M^{\mu}$ and $\mathcal{N}_N^{\nu}$ respectively and the notation in table 4 applies with $\mathcal{T} = \mathcal{S}$. Following (3.2) and (3.3), the two models are *de facto* endowed with joint and predictive probabilities: (Π, P_S) for $\mathcal{M}_M^{\mu}$ and (χ, Q_S) for $\mathcal{N}_N^{\nu}$ respectively. In the rest of this section, we proceed under the convention that the reference probability is the predictive probability P_S associated with $\mathcal{M}_M^{\mu}$.

Definition 4.1 raises technical issues with the derivation of $\Delta_{\mathcal{B}}^{A}$ when A and B are not 'distinct', e.g. when they include common parameters. A Bayesian reformulation of definition 4.1 which implicitly addresses these technicalities runs as follows. Let Θ be a parameter space such that $A \subset \Theta$ and $B \subset \Theta$. Let Θ be endowed with the σ-field $\mathcal{A} \vee \mathcal{B}$, defined as the smallest σ-field generated by $\mathcal{A} \cup \mathcal{B}$.

Theorem 4.1. *$\mathcal{M}_M^{\mu}$ exactly encompasses $\mathcal{N}_N^{\nu}$ if and only if there exists a probability Π^* on $\{\Theta \times \mathcal{S}, (\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}\}$ such that:*
(i) $\forall G \in \mathcal{A}, \forall X \in \mathcal{S}, \Pi^*(G \times X) = \Pi(G \times X)$;
(ii) $N_{\mathcal{B}}^{S} = N_{\mathcal{B}}^{*S}$, where $N_{\mathcal{B}}^{*S}$ is the posterior transition derived from Π^*;
(iii) $\mathcal{B} \perp\!\!\!\perp \mathcal{S} | \mathcal{A}$ under Π^*.

Proof : See Appendix. ■

If in particular $\mathcal{B} \subset \mathcal{A}$ and χ is the restriction of Π to $\mathcal{B} \otimes \mathcal{S}$, then $\mathcal{N}$ is derived from $\mathcal{M}$ by marginalization and conditions (i)-(iii) are satisfied with $\Pi^* = \Pi$. This result formalizes the heuristic claim that if $\mathcal{N}_N^{\nu}$ is explicitly 'nested' within $\mathcal{M}_M^{\mu}$, then it ought to be encompassed by the latter.

Under the conditions of theorem 4.1, the two inferential models are nested within a 'super-model' characterized by Π^* though they are not treated symmetrically. In particular, the restriction of Π^* to $\mathcal{B} \otimes \mathcal{S}$ need not coincide with χ and, hence Q_S cannot be retrieved from Π^*. In a number of contexts, such as that of model choice and the analysis in Florens and Scotto [15], it may be desirable to treat the two models symmetrically (except for the encompassing condition itself which is inherently asymmetric). This is achieved by indexing the two models and treating the index i as an additional parameter. Let $I = \{1, 2\}$ and $\mathcal{I} = \mathcal{P}(I)$. Let Ψ denote a probability on $\{I \times \Theta \times \mathcal{S}, \mathcal{I} \otimes (\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}\}$. The necessary additional notation is:

(i) α for the marginal probability of Ψ on $(I, \mathcal{I})$;
(ii) $\Psi^1_{\mathcal{A} \otimes \mathcal{S}}$ for the restriction of Ψ to $\mathcal{A} \otimes \mathcal{S}$, conditionally on $i = 1$;
(iii) $\Psi^2_{\mathcal{B} \otimes \mathcal{S}}$ for the restriction of Ψ to $\mathcal{B} \otimes \mathcal{S}$, conditionally on $i = 2$.

Theorem 4.2. *$\mathcal{M}^{\mu}_{M}$ exactly encompasses $\mathcal{N}^{\nu}_{N}$ if and only if there exists a probability Π on $\{I \times \Theta \times S, \mathcal{I} \otimes (\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}\}$ such that:*
(i) $\alpha(i) > 0$ for $i = 1, 2$;
(ii) $\Psi^{1}_{\mathcal{A}\otimes\mathcal{S}} = \Pi$;
(iii) $\Psi^{2}_{\mathcal{B}\otimes\mathcal{S}} = \chi$;
(iv) $\mathcal{B} \perp\!\!\!\perp \mathcal{S} | \mathcal{A}$ conditionally on $i = 1$ and $\mathcal{A} \perp\!\!\!\perp \mathcal{S} | \mathcal{B}$ conditionally on $i = 2$;
(v) $\mathcal{B} \perp\!\!\!\perp \mathcal{I} | \mathcal{S}$ under Ψ. ■

Proof: See Appendix.

Condition (iv) simply states that $\mathcal{A}$ and $\mathcal{B}$ are 'sufficient' parameterizations within their respective models. The nesting in theorem 4.2 is partially arbitrary and hence is not unique. Generally, the αs can be arbitrarily chosen as well as the transition $\mathcal{B} \times \mathcal{A}$ which is implicit in the construction of a probability on $(\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}$, conditionally on $i = 2$. Nevertheless, theorem 4.2 provides a formulation which is convenient when the index i is itself a parameter of interest, as in the literature on model choice. The relationship between encompassing and model choice is discussed in section 5.5.

Bayesian exact encompassing relies upon the existence of a transition between the posterior probabilities. An intriguing issue is whether or not it also implies the existence of a transition between the sampling probabilities. A general answer to that question is provided by the next theorem.

Theorem 4.3. *Let $\mathcal{M}^{\mu}_{M}$ and $\mathcal{N}^{\nu}_{N}$ be two Bayesian inferential models with equivalent predictive probabilities. Let ρ denote a probability on $\mathcal{A} \vee \mathcal{B}$ such that $\rho_{\mathcal{A}} = \mu_{\mathcal{A}}$ and $\rho_{\mathcal{B}}$ is equivalent to $\nu_{\mathcal{B}}$. Let $\Delta^{\mathcal{A}}_{\mathcal{B}}$ and $K^{\mathcal{S}}_{\mathcal{B}}$ denote the corresponding conditional transition probabilities. The following two conditions are equivalent:*
(i) ρ is such that $\mathcal{M}^{\mu}_{M}$ exactly encompasses $\mathcal{N}^{\nu}_{N}$ with transition $\Delta^{\mathcal{A}}_{\mathcal{B}}$;
(ii) ρ is such that:

$$\forall X \in \mathcal{S}, \quad \int_{A} P^{\mathcal{A}}_{\mathcal{S}}(X) dK^{\mathcal{B}}_{\mathcal{A}} = \left(\frac{d\nu_{\mathcal{B}}}{d\rho_{\mathcal{B}}}\right) \int_{X} \left(\frac{dP_{\mathcal{S}}}{dQ_{\mathcal{S}}}\right) dQ^{\mathcal{B}}_{\mathcal{S}} \tag{4.12}$$

■

Proof: See Appendix.

It follows from theorem 4.3 that exact encompassing does not entail the existence of a transition on $\mathcal{B} \times \mathcal{A}$ that can be used to directly transform $P^{\mathcal{A}}_{\mathcal{S}}$ into $Q^{\mathcal{B}}_{\mathcal{S}}$ unless additional conditions are imposed on the prior and predictive probabilities. This is the object of the concepts of coherent and strong (exact) encompassing which are introduced below.

Definition 4.4. *$\mathcal{M}^{\mu}_{M}$ coherently (exactly) encompasses $\mathcal{N}^{\nu}_{N}$ if and only if:*
(i) $\mathcal{M}^{\nu}_{M}$ exactly encompasses $\mathcal{N}^{\nu}_{N}$ with pseudo-true value $\Delta^{\mathcal{A}}_{\mathcal{B}}$;
(ii) $\mu_{\mathcal{A}}$ and $\nu_{\mathcal{B}}$ are coherent with each other relative to $\Delta^{\mathcal{A}}_{\mathcal{B}}$ in the sense that:

$$\forall F \in \mathcal{B}, \quad \nu_{\mathcal{B}}(F) = \int_{A} \Delta^{\mathcal{A}}_{\mathcal{B}}(F) d\mu_{\mathcal{A}} \tag{4.13}$$

Condition (4.13) entails that $\nu_{\mathcal{B}}$ coincides with $\rho_{\mathcal{B}}$, as defined in theorem 4.3 and, hence, that $d\nu_{\mathcal{B}}/d\rho_{\mathcal{B}} = 1$. Under the conditions of theorem 4.2, it is reformulated as:
(ii)' $\mathcal{B} \perp\!\!\!\perp \mathcal{I}$ under Ψ.

Definition 4.5. *$\mathcal{M}_M^\mu$ strongly (exactly) encompasses $\mathcal{N}_N^\nu$ if and only if:*
(i) $\mathcal{M}_M^\mu$ exactly encompasses $\mathcal{N}_N^\nu$;
(ii) $\forall X \in \mathcal{S}, P_{\mathcal{S}}(X) = Q_{\mathcal{S}}(X)$.

Under the conditions of theorem 4.2, condition (ii) is reformulated as:
(ii)' $\mathcal{S} \perp\!\!\!\perp \mathcal{I}$ under Ψ.

Theorem 4.4. *Strong encompassing implies coherent encompassing.*
Proof: The proof is immediate under the conditions of theorem 4.2 since:

$$\mathcal{B} \perp\!\!\!\perp \mathcal{I} \mid \mathcal{S} \quad \text{and} \quad \mathcal{S} \perp\!\!\!\perp \mathcal{I} \Rightarrow \mathcal{B} \perp\!\!\!\perp \mathcal{I} \qquad \text{under } \Psi$$

■

It also follows from theorems 4.2 and 4.4 that strong encompassing can be reformulated in terms of the existence of a transition probability between sampling probabilities.

Theorem 4.5. *$\mathcal{M}_M^\mu$ strongly encompasses $\mathcal{N}_N^\nu$ if and only if there exists a transition probability $K_{\mathcal{A}}^{\mathcal{B}}$ such that:*

$$(i)\ \forall X \in \mathcal{S}, \quad Q_{\mathcal{S}}^{\mathcal{B}}(X) = \int_A P_{\mathcal{S}}^{\mathcal{A}}(X) dK_{\mathcal{A}}^{\mathcal{B}} \tag{4.14}$$

$$(ii)\ \forall E \in \mathcal{A}, \quad \mu_{\mathcal{A}}(E) = \int_B K_{\mathcal{A}}^{\mathcal{B}}(X) d\nu_{\mathcal{B}} \tag{4.15}$$

The pseudo-true value $\Delta_{\mathcal{B}}^{\mathcal{A}}$ is derived from the joint probability ρ on $\mathcal{A} \otimes \mathcal{B}$ associated with the pair $(\nu_{\mathcal{B}}, K_{\mathcal{A}}^{\mathcal{B}})$.
Proof: See Appendix. ■

In concluding this section, we emphasize that the concepts of coherent and strong encompassing differ fundamentally in their treatment of the predictive probabilities and so will be used in different contexts. Coherent encompassing is relevant in situations where one wishes to compare models under a common body of prior knowledge. We should nevertheless not rule out the possibility that the models could also be compared under mutually incoherent prior probabilities, e.g. as initially specified by their respective builders, since the specifications of a sampling model and a prior probability are typically interrelated. Strong encompassing is relevant within such contexts as that of a 'hierarchical' (joint) model where a transition $K_{\mathcal{B}}^{\mathcal{A}}$ is explicitly introduced to reduce the dimensionality of a parameter space $\mathcal{A}$ and where (4.15) then states the coherency condition to be satisfied by the corresponding hierarchical prior probability.

Example 4.4: Consider the hierarchical models:
$P_{\mathcal{S}}^{\mathcal{A}}:\quad \mathrm{s}|\mathbf{a} \sim \mathsf{N}_T\left(\mathbf{a}, \sigma^2\mathbf{I}_T\right), \quad \mathbf{a} \in \mathbb{R}^T;$
$K_{\mathcal{A}}^{\mathcal{B}}:\quad \mathbf{a}|b \sim \mathsf{N}_1\left(b \cdot \boldsymbol{\iota}, \tau^2\mathbf{I}_T\right), \quad b \in \mathbb{R}, \quad \boldsymbol{\iota}' = (1 \ldots 1);$
$Q_{\mathcal{S}}^{\mathcal{B}}:\quad \mathrm{s}|b \sim \mathsf{N}_T\left(b \cdot \boldsymbol{\iota}, (\sigma^2 + \tau^2)\,\mathbf{I}_T\right)$
whence $\mathcal{M}_M^{\mu}$ strongly encompasses $\mathcal{N}_N^{\nu}$ under the coherent priors:

$$\nu_{\mathcal{B}} : b \sim \mathsf{N}_1\left(b_0, \nu_0\right) \text{ and } \mu_{\mathcal{A}} : \mathbf{a} \sim \mathsf{N}_T\left(b_0 \cdot \boldsymbol{\iota}, \tau^2\mathbf{I}_T + \nu_0\boldsymbol{\iota}\boldsymbol{\iota}'\right).$$

■

5 Specificity

Exact encompassing provides conditions under which the functions in $[\mathcal{B}]_\infty$ can be 'estimated' equivalently either directly within $\mathcal{N}_N$ or indirectly within $\mathcal{M}_M$ via the transition probability $\Delta_{\mathcal{B}}^{\mathcal{A}}$. In practice, however, it will often be the case that the two procedures yield different solutions, especially when $\mathcal{M}_M$ and $\mathcal{N}_N$ have initially been designed for different purposes or by different investigators. Exactly as Lecam [26] uses a notion of deficiency to measure a lack of sufficiency, we introduce two concepts of 'specificity' aimed at measuring a lack of exact encompassing.

5.1 p-specificity

If attention is restricted to a specific random variable $g \in [\mathcal{B}]_\infty$, then its 'estimators' $N_{\mathcal{B}}^{\mathcal{S}}(g)$ and $\left(M_{\mathcal{A}}^{\mathcal{S}} \circ \Delta_{\mathcal{B}}^{\mathcal{A}}\right)(g)$ can be compared by means of an L_P-norm on $(S, \mathcal{S})$ endowed with a reference probability $P_{\mathcal{S}}^0$.

Definition 5.1. *The p-specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$ with respect to g, given $P_{\mathcal{S}}^0$ and a class of transition probabilities $\mathcal{D}$ is:*

$$\sigma_g\left(\mathcal{N}_N; \mathcal{M}_M\right) = \inf_{\Delta_{\mathcal{B}}^{\mathcal{A}} \in \mathcal{D}} \left\| N_{\mathcal{B}}^{\mathcal{S}}(g) - \left(M_{\mathcal{A}}^{\mathcal{S}} \circ \Delta_{\mathcal{B}}^{\mathcal{A}}\right)(g) \right\|_p \tag{5.1}$$

This definition can be extended to $[\mathcal{B}]_\infty$.

Definition 5.2. *The p-specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$, given $P_{\mathcal{S}}^0$ and $\mathcal{D}$, is:*

$$\sigma\left(\mathcal{N}_N; \mathcal{M}_M\right) = \sup_{g \in [\mathcal{B}]_\infty} \sigma_g\left(\mathcal{N}_N; \mathcal{M}_M\right) \quad \textit{subject to} \quad \|g\|_\infty \leq 1 \tag{5.2}$$

While the second definition is conceptually interesting, and is related by duality to the concept of deficiency in Lecam [26], it will prove impossible to evaluate except for trivial cases. Definition 5.1 is more operational under suitable choices of g. An example is provided in section 5.3.

5.2 φ-specificity

Under suitable existence conditions, we can also analyze the expectation of a measure of the 'divergence' between the two estimation procedures $N^S_{\mathcal{B}}$ and $M^S_{\mathcal{A}} \circ \Delta^{\mathcal{A}}_{\mathcal{B}}$. The concept of φ-divergence, as discussed e.g. in Cziszar [7] - also see Florens and Scotto [15] for Bayesian applications - is convenient for that purpose. Let φ be a real valued convex function defined on $\mathbb{R}_+$ such that $\varphi(1) = 0$. The φ-divergence between two probabilities P and Q is defined as:

$$D_\varphi (P; Q) = \int \varphi \left(\frac{dP}{dQ} \right) dQ \tag{5.3}$$

Special cases of φ-divergence are characterized as follows: (i) The negative entropy D_E (also called the Kullback-Leibler comparison): $\varphi(x) = x \log x$; (ii) The square of the Hellinger distance $D_H : \varphi(x) = (\sqrt{x} - 1)^2$; (iii) The total variation distance $D_1 : \varphi(x) = \frac{1}{2} |x - 1|$; (iv) The χ^2 comparison $D_2 : \varphi(x) = (x - 1)^2$. Except for D_1, these measures are not 'distances' in the strict sense.

Definition 5.3. *The φ-specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$, given $P^0_{\mathcal{S}}$ and $\mathcal{D}$, is:*

$$\tau_\varphi (\mathcal{N}_N; \mathcal{M}_M) = \min_{\Delta^{\mathcal{A}}_{\mathcal{B}} \in D} \int D_\varphi \left(N^S_{\mathcal{B}}; M^S_{\mathcal{A}} \circ \Delta^{\mathcal{A}}_{\mathcal{B}} \right) dP^0_{\mathcal{S}} \tag{5.4}$$

This definition of the φ-specificity raises no conceptual problems within a Bayesian framework when $N^S_{\mathcal{B}}$ and $M^S_{\mathcal{A}}$ are posterior probabilities. It cannot be applied to mutually singular classical estimation procedures, such as Dirac measures, since the φ-divergence is then maximal independently of the choice of a transition. Other definitions of specificity based, for example, on the notion of weak convergence distance in Billingsley [4] do not suffer from that limitation but are not as operational as that of φ-specificity.

5.3 A special case

The search for a solution to (5.4) is often complicated. We next discuss a special case for which an explicit solution is available and which generalizes the approach followed by Florens *et al.* [10].

Let $\mathcal{A} = \{a_0\}$. $\mathcal{M}$ then consists of a single probability $P^0_{\mathcal{S}}$ which also serves as the reference probability. The estimation procedure associated with $\mathcal{M}_M$ is a Dirac $D^S_{\mathcal{A},\hat{a}}$ with $\hat{a}(s) = a_0, P^0_{\mathcal{S}}$-almost surely. It follows that $\Delta^{\mathcal{A}}_{\mathcal{B}}$ simplifies to a single conditional probability on $\mathcal{B}$ given a_0. We assume that both $\Delta^{\mathcal{A}}_{\mathcal{B}}$ and $N^S_{\mathcal{B}}$ are dominated probabilities with respective densities $\delta(b)$ and $\nu(b|s)$ relative to a measure on $(B, \mathcal{B})$ with differential element db. $\mathcal{D}$ is defined as the set of all such $\delta(\cdot)$s. This special case is of interest for two main reasons:

(i) it can serve as the basis for a test of the *single* hypothesis that a model $\mathcal{M}_M$ with $A = \{a_0\}$ encompasses a rival $\mathcal{N}_N$;

(ii) as discussed below, it is instrumental in the pointwise construction of a transition probability $\Delta_{\mathcal{B}}^{\mathcal{A}}$ which constitutes an operational alternative to a solution of (5.4) when $\mathcal{A}$ is not a singleton.

Under our simplifying assumptions and using the negative entropy D_E, the φ-specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$ is:

$$\begin{aligned} \tau_\varphi(\mathcal{N}_N; \mathcal{M}_M) &= \min_{\delta \in \mathcal{D}} \int_S \left\{ \int_{\mathcal{B}} \log\left[\tfrac{\delta(b)}{\nu(b|s)}\right] \delta(b)\, db \right\} dP_S^0 \\ &= \min_{\delta \in \mathcal{D}} \left\{ \int_{\mathcal{B}} \log\left[\frac{\delta(b)}{\delta_E(b)}\right] \delta(b)\, db - \log K \right\} \end{aligned} \tag{5.5}$$

where δ_E is the auxiliary density:

$$\delta_E(b) = K^{-1} \cdot \exp\left[\int_S \log \nu(b|s) \cdot dP_S^0\right] \tag{5.6}$$

and K is its integrating constant which is shown to be less than one by application of Jensen's inequality. Note that K is unaffected by the choice of $\delta(\cdot)$. Hence the optimal solution in (5.5) is given by δ_E itself and τ_φ equals $-\log K$. If, furthermore, $\nu(b|s)$ is a posterior density derived from a prior density $\nu(b)$, then:

$$\delta_E(b) \propto \nu(b) \cdot \exp\left[\int_S \log q(s|b) \cdot dP_S^0\right] \tag{5.7}$$

The transition δ_E often is easier to evaluate than a solution to (5.4). Consider, for example the case where $\mathcal{M}$(unit root) and $\mathcal{N}$ are characterized as follows:

$$\mathcal{M} \;:\; s_i = s_{i-1} + u_i \;; \quad u_i \sim \mathsf{IN}(0, 1) \tag{5.8}$$

$$\mathcal{N} \;:\; s_i = b s_{i-1} + v_i \;; \quad v_i \sim \mathsf{IN}(0, 1) \tag{5.9}$$

Florens *et al.* [10] demonstrate that, if $\nu(b) \propto 1$, then δ_E is the density associated with the normal distribution $\mathsf{N}\left(1, \frac{2}{n(n-1)}\right)$, where n is the sample size.

Using the Hellinger distance D_H instead of D_E, the optimal transition δ_H is given by:

$$\delta_H(b) \propto \left\{\int_S [\nu(b|s)]^{\frac{1}{2}}\, dP_S^0\right\}^2 \tag{5.10}$$

As suggested earlier, the special case just discussed is instrumental in the derivation of operational - though non-optimal choices - for $\Delta_{\mathcal{B}}^{\mathcal{A}}$ in the more general case where $\mathcal{A}$ is not a singleton. They are defined as transition probabilities which, to every

a in $\mathcal{A}$, associate a transition which is optimal for that specific a. For example, a $\delta_E(b|a)$ which follows from (5.6) is:

$$\delta_E(b|a) \propto \exp\left[\int_S \log \nu(b|s) \cdot dP_S^A\right] \tag{5.11}$$

In line with the discussion in section 6 below, $\delta_E(b|a)$ ought to have the same asymptotic properties as an optimal transition, whence follows its usefulness as an operational alternative to the latter. An explicit comparison is provided in section 8 in the context of the choice of regressors problem.

5.4 Bayesian encompassing tests

We can usefully draw a parallel between the construction of classical and Bayesian encompassing test procedures.

Classical test procedures are based on a pseudo-true value $b(a)$, which is usually defined as the plim under P_S^A of an estimator $\hat{b}$ of b so that it typically minimizes the Kullback-Leibler divergence between the two sampling distributions. A distance between $\hat{b}$ and $b(\hat{a})$, where $\hat{a}$ is an estimator of a, is then evaluated and calibrated in accordance with conventional test principles (Wald or score). See Mizon and Richard [28] for details and also Gourieroux *et al.* [18] and [17].

A similar approach applies within a Bayesian framework following testing principles described e.g. in Florens and Mouchart [12]. As discussed above, a Bayesian pseudo-true value $\Delta_{\mathcal{B}}^{\mathcal{A}}$ is a transition probability which minimizes the specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$, where specificity is defined as the $\mathcal{M}$-predictive expectation of the divergence between the $\mathcal{N}$- and $\mathcal{M}$- posterior distributions of b. The divergence itself, evaluated under the optimal transition, is a function of the actual sample and may, therefore, serve as an encompassing test statistic. The two alternative models under consideration are characterized by their predictive densities.

For ease of presentation, we restrict attention to the φ-specificity, as defined in section 5.2, assuming that there is a unique solution to the optimization problem (5.4). The statistic of interest is then given by:

$$\xi(s) = D_\varphi\left(N_{\mathcal{B}}^S; M_{\mathcal{A}}^S \circ \Delta_{\mathcal{B}}^{\mathcal{A}}\right) \tag{5.12}$$

where $\xi(s)$ also depends on $\left(M_{\mathcal{A}}^S, N_{\mathcal{B}}^S, \varphi\right)$, but such arguments are omitted for ease of notation. The M_1-predictive expectation of $\xi(s)$ is the specificity itself, as given in (5.4). The predictive distribution of $\xi(s)$, under either $\mathcal{M}$ or $\mathcal{N}$, often is analytically intractable but can be evaluated by means of a conceptually straightforward Monte Carlo simulation. An application of this principle is discussed in section 8.

5.5 Encompassing and Model Choice

The object of encompassing is not that of choosing between two inferential models $\mathcal{M}_M$ and $\mathcal{N}_N$. It is instead that of examining whether or not a rival model $\mathcal{N}_N$ is *redundant* relative to one's preferred model $\mathcal{M}_M$, for the purpose of conducting inference on b. The finding that $\mathcal{M}_M$ does not encompass $\mathcal{N}_N$ indicates that the latter contains information *relative to b* that cannot be retrieved from the former. It does not imply that $\mathcal{N}_N$ is to be preferred to or chosen against $\mathcal{M}_M$ since, in particular, the parameters of $\mathcal{N}$ may not even be of direct interest to the proprietor of $\mathcal{M}$. The object of the exercise is that of validating $\mathcal{M}_M$ in light of 'fresh' evidence provided by $\mathcal{N}_N$ and lack of encompassing typically leads to further improvements of $\mathcal{M}_M$ itself. See Hendry and Richard [21] for further discussion of the role of encompassing as a key component of a progressive modelling strategy.

This being said, we can usefully examine the relationship between encompassing test statistics and Bayes factors for the pair $(\mathcal{M}_M, \mathcal{N}_N)$. Depending on which measure of divergence is being used, the relationship can be rather muddled though an interesting comparison emerges if we restrict attention to the encompassing test statistic which is based on negative entropy. Assuming that the relevant distributions can be characterized by (well behaved) density functions, the negative entropy encompassing test statistic is given by:

$$\xi(s) = \int_B \log\left(\frac{\nu^*(b|s)}{\nu(b|s)}\right) \cdot \nu^*(b|s)db \tag{5.13}$$

A rearrangement of factors leads to the expression:

$$\xi(s) = -\log\left(\frac{p(s)}{q(s)}\right) + \int_B \log\left(\frac{\pi^*(b,s)}{\chi(b,s)}\right) \cdot \nu^*(b|s)db \tag{5.14}$$

where the notation is in line with that in table 3. Hence $\xi(s)$ is given by the difference between the $\mathcal{M}$-posterior expectation of the log-ratio of the joint probabilities on (b,s) corresponding to both models, and the log of the corresponding Bayes factors. In line with the general result derived in theorem 4.5, the $\mathcal{M}$-posterior density $\nu^*(b|s)$, as defined in (2.9), may be rewritten as:

$$\nu^*(b|s) = \frac{\tilde{\nu}(b) \cdot \tilde{q}(s|b)}{p(s)} \tag{5.15}$$

where

$$\tilde{\nu}(b) = \int_A \mu(a) \cdot \delta(b|a)da \tag{5.16}$$

$$\tilde{q}(s|b) = \int p(s|a) \cdot \kappa(a|b)da \tag{5.17}$$

$$\kappa(a|b) = \frac{\mu(a) \cdot \delta(b|a)}{\tilde{\nu}(b)} \tag{5.18}$$

Hence $\tilde{\nu}(b)$ and $\tilde{q}(s|b)$ are $\mathcal{M}$-coherent prior and sampling densities which could be used to derive $\nu^*(b|s)$ directly by application of Bayes theorem, as in (5.15) noting that $p(s) = \int_B \tilde{q}(s|b)\tilde{\nu}(b)db$. It follows that (5.14) may be rewritten as:

$$\xi(s) = -\log\left(\frac{p(s)}{q(s)}\right) + \int_B \left[\log\left(\frac{\tilde{q}(s|b)}{q(s|b)}\right) + \log\left(\frac{\tilde{\nu}(b)}{\nu(b)}\right)\right] \nu^*(b|s)db \qquad (5.19)$$

If the specificity itself is evaluated under a coherent prior, then (5.19) simplifies to:

$$\tilde{\xi}(s) = -\log\left(\frac{p(s)}{\tilde{q}(s)}\right) + \int_B \log\left(\frac{\tilde{q}(s|b)}{q(s|b)}\right) \nu^*(b|s)db \qquad (5.20)$$

where $\tilde{q}(s) = \int_B q(s|b)\tilde{\nu}(b)db$. High posterior odds in favor of $\mathcal{M}$ contribute to lowering the value of the encompassing test statistic $\xi(s)$ - and conversely - but the latter also depends on additional terms, as given in (5.19) and (5.20).

Finally it must be emphasized that, in much of the literature on model choice, the default option is one of prior *independence* between a and b. Such an assumption is inconsistent with the fact that both models are meant to capture a common sampling process. Further, it is instrumental in generating some of paradoxes that plague applied work in the form of extreme values for the posterior odds. See e.g. the discussion in Kiefer and Richard [24]. In contrast, encompassing explicitly requires stochastic dependence between a and b, in the form either of a genuine prior transition $\delta(b|a)$ or of a transition which minimizes the specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$. Formula (5.20) appears to be of special interest in the (common) situation where $\mathcal{M}$ is one's 'preferred' model. In such cases there would be little interest in rejecting $\mathcal{M}$ in favor of a rival model $\mathcal{N}$ whose specificity relative to $\mathcal{M}$ is small. Neither should rejection result from the use of a prior $\nu(b)$ which is not coherent with the $\mathcal{M}$-prior density $\mu(a)$. Formula (5.20) suggests evaluating posterior odds under an $\mathcal{N}$-prior $\tilde{\nu}(b)$ which (i) is coherent with $\mu(a)$ in the sense of (5.16), and (ii) minimizes the specificity of $\mathcal{N}$ relative to $\mathcal{M}$. Such a prior constitutes a 'fixed-point' prior in the sense of (5.4) and (5.16). Its existence in general is yet an open issue but approximate solutions can be evaluated in line with our discussion in section 7 below.

6 Sequential Encompassing and Asymptotics

In previous sections, we implicitly adopted a 'global' mode of analysis, specifically restricting attention to the derivation of a single transition probability relative to a sample space $(S, \mathcal{S})$ of fixed dimensionality. However, a broad class of statistical problems require a 'sequential' mode of reasoning relative to a sequence of embedded sample spaces $\{(S_n, \mathcal{S}_n)\}$. Consider, in particular, the problem of analyzing the limiting (asymptotic) behavior of a sequence of encompassing transition probabilities. The 'asymptotic' experiments are denoted by $\mathcal{E} = \{A \times S, \mathcal{A} \vee \mathcal{S}, \Pi\}$ and $\mathcal{F} = \{B \times S, \mathcal{B} \vee \mathcal{S}, \chi\}$ and $\mathcal{S}$ is endowed with a filtration $\mathcal{S}_n \rightarrow \mathcal{S}$. Let Π_n denote the restriction of Π to $\mathcal{A} \vee \mathcal{S}_n$ and $\mathcal{E}_n = \{A \times S, \mathcal{A} \vee \mathcal{S}_n, \Pi_n\}$. A similar notation applies to $\mathcal{F}$ and we have $\mathcal{F}_n = \{B \times S, \mathcal{B} \vee \mathcal{S}_n, \chi_n\}$. Finally, let $\left(\Delta^{\mathcal{A}}_{\mathcal{B}}\right)_n$ denote the encompassing transition associated with the sample size n (i.e. the transition which minimizes the specificity of $\mathcal{F}_n$ relative to $\mathcal{E}_n$ under the working criterion). In particular, $\left(\Delta^{\mathcal{A}}_{\mathcal{B}}\right)_n$ extends Π_n to a probability Π^*_n on $(\mathcal{A} \vee \mathcal{B}) \vee \mathcal{S}_n$.

An heuristic argument runs as follows: 'well-behaved' inferential procedures should converge towards limiting Dirac distributions which, under P^A_S in particular, would be centered around a and $b(a)$ respectively, where $b(a)$ is a (classical) pseudo-true value of b. In such a case, sequences of optimal encompassing transitions $\left(\Delta^{\mathcal{A}}_{\mathcal{B}}\right)_n$ would themselves converge towards a Dirac distribution centered on $b(a)$ and the specificity of $\mathcal{F}_n$ relative to $\mathcal{E}_n$ would tend to zero (under P^A_S). A formal proof is found in Florens and Richard [14] for the case where the parameter spaces A and B are discrete (covering situations where, as discussed in Berk [2], [3], the support of the limiting distribution on $\mathcal{B}$ is a singleton). Attempts to generalize this result have yet to deal with the difficulty that the sequence $\{\Pi^*_n\}$ does *not* constitute a projective system, in the sense that the restriction of Π^*_n on $(\mathcal{A} \vee \mathcal{B}) \vee \mathcal{S}'_n$ for $n' < n$ generally does not coincide with $\Pi^*_{n'}$. It follows for example, that we might have exact encompassing for all ns:

$$N^{\mathcal{S}_n}_{\mathcal{B}} = M^{\mathcal{S}_n}_{\mathcal{A}} \circ \left(\Delta^{\mathcal{A}}_{\mathcal{B}}\right)_n \quad \forall n \tag{6.1}$$

and yet:[9]

$$N^{\mathcal{S}_{n'}}_{\mathcal{B}} \neq M^{\mathcal{S}_{n'}}_{\mathcal{A}} \circ \left(\Delta^{\mathcal{A}}_{\mathcal{B}}\right)_n \quad \text{for } n' < n \tag{6.2}$$

Further, in the context of (dynamic) sequential models it would be natural to condition encompassing transitions on a σ-field $\mathcal{T}_n$ of 'exogenous' variables - as in section 8 - and possibly also on the σ-field $\mathcal{S}_{n-1}$ of 'lagged endogenous' variables. The probability Π_n then has to be extended to a probability Π^*_n on $(\mathcal{A} \vee \mathcal{B}) \vee \mathcal{S}_n | \mathcal{T}_n \vee \mathcal{S}_{n-1}$ by means of a transition $\left(\Delta^{\mathcal{A} \vee \mathcal{T}_n \vee \mathcal{S}_{n-1}}\right)_n$. Here again the sequence $\{\Pi^*_n\}$ does not constitute a projective system.

We shall not discuss sequential encompassing further and, for the rest of the paper, simply assume that both $\left\{N^{\mathcal{S}_n}_{\mathcal{B}}\right\}$ as well as sequences $\left(\Delta^{\mathcal{A}}_{\mathcal{B}}\right)_n$ of optimal transitions converge towards a limiting distribution $D^{\mathcal{A}}_{\mathcal{B}}$. In fact, as discussed next, convergence towards that common limiting distribution will be used as an

[9] Note that $\left(\Delta^{\mathcal{A}}_{\mathcal{B}}\right)_n$ is an admissible transition for $n' < n$, since under Π^*_n if $\mathcal{B} \perp\!\!\!\perp \mathcal{S}_n | \mathcal{A}$, then $\mathcal{B} \perp\!\!\!\perp \mathcal{S}_{n'} | \mathcal{A}$.

important criterion in the selection of operational 'approximations' to optimal transitions that are intractable.

7 Approximate encompassing

There are few cases where there exist operational solutions to the minimization problems in definitions 5.1 to 5.3. One such case has been discussed in section 5.3 above. Another case (discussed in Florens and Richard, 1989) is where the sample spaces A and B are finite. Numerical solutions might also exist though a thorough discussion of their implementation goes beyond the objective of this paper. We limit ourselves here to three approximate solutions which, in addition to the one already discussed in section 5.3, could usefully be implemented for a broad class of problems. In line with the discussion in section 6, our analysis will be largely heuristic as far as asymptotic properties are concerned.

7.1 A 'Marginalized Likelihood' pseudo-true value

The first approximation we propose is a straightforward generalization of the classical notion of pseudo-true value which is defined as the plim under $P_{\mathcal{S}}^{A}$ of a point estimate $N_{\mathcal{B}}^{\mathcal{S}}$. It consists of marginalizing the inferential procedure $N_{\mathcal{B}}^{\mathcal{S}}$ with respect to $\mathcal{S}$ using the sampling distribution $P_{\mathcal{S}}^{A}$. Hence, let:

$$\tilde{\Delta}_{\mathcal{B}}^{A} = P_{\mathcal{S}}^{A} \circ N_{\mathcal{B}}^{\mathcal{S}} \tag{7.1}$$

The transition $\tilde{\Delta}_{\mathcal{B}}^{A}$ is clearly not optimal in the sense of definitions 5.1 to 5.3 since, in particular, if the two estimation procedures under consideration did coincide with each other, then the optimal transition would be the Dirac distribution associated with the identity mapping, whereas $\tilde{\Delta}_{\mathcal{B}}^{A}$ in (7.1) would not. Beyond considerations of computational convenience, the following asymptotic theorem provides the rationale for using $\tilde{\Delta}_{\mathcal{B}}^{A}$ as an approximation to the optimal $\Delta_{\mathcal{B}}^{A}$ when the latter is not available (For notational convenience, transitions are not indexed by sample size as they were in section 6).

Theorem 7.1. *If:*
(i) $\mathcal{M}_M$ *is exactly estimable;*[10]
(ii) there exists a transition $\mathcal{D}_{\mathcal{B}}^{A}$ *such that:*

$$\forall F \in \mathcal{B}, \quad \nu_{\mathcal{B}}^{\mathcal{S}}(F) \to D_{\mathcal{B}}^{A}(F), \quad \Pi\text{- } \textit{almost surely} \ ,$$

then:

$$\forall F \in \mathcal{B}, \quad \tilde{\Delta}_{\mathcal{B}}^{A}(F) \to D_{\mathcal{B}}^{A}(F), \quad \mu\text{- } \textit{almost surely} \ ,$$

[10] In the sense that the posterior expectation of any integrable function of the parameters converges towards that function Π-almost surely.

Proof: If we factorize Π into the product of $P_{\mathcal{S}}^{\mathcal{A}}$ and $\mu_{\mathcal{A}}$, then condition (ii) implies that:

$$\left\{\nu_{\mathcal{B}}^{\mathcal{S}}(F) \to D_{\mathcal{B}}^{\mathcal{A}}(F), P_{\mathcal{S}}^{\mathcal{A}}\text{- a.s. }\right\}\mu_{\mathcal{A}}\text{- a.s.}$$

Further, $\nu_{\mathcal{B}}^{\mathcal{S}}(F)$ is bounded above by 1. Hence, by Lebesgue's theorem (see Dellacherie and Meyer, 1975 Chap. II):

$$\left\{\tilde{\Delta}_{\mathcal{B}}^{\mathcal{A}} = E\left(\nu_{\mathcal{B}}^{\mathcal{S}}(F) \mid \mathcal{A}\right) \to E\left(D_{\mathcal{B}}^{\mathcal{A}}(F) \mid \mathcal{A}\right) = D_{\mathcal{B}}^{\mathcal{A}}(F)\right\}\mu_{\mathcal{A}}\text{- a.s.}$$

■

Explicit comparisons between $\tilde{\Delta}_{\mathcal{B}}^{\mathcal{A}}$ and the optimal transition are found in Florens and Richard [14] for the case where the parameter spaces are finite and in section 8 for the choice of regressors problem. When the baseline model is a Bayesian inferential model $\mathcal{M}_M^{\mu}$, then (7.1) implicitly defines an auxiliary joint distribution $\tilde{\Pi}$ on $\mathcal{A} \otimes \mathcal{B}|\mathcal{S}$ as the product of the transition probabilities $\mu_{\mathcal{A}}^{\mathcal{S}}$ and $\nu_{\mathcal{B}}^{\mathcal{S}}$ (so that $\mathcal{A} \perp\!\!\!\perp \mathcal{B}|\mathcal{S}$ under $\tilde{\Pi}$) which is then marginalized with respect to $\mathcal{S}$ under the predictive $P_{\mathcal{S}}$ and finally conditionalized on $\mathcal{A}$. In terms of density functions:

$$\tilde{\delta}(b|a) = \frac{1}{\mu(a)} \int \mu(a|s)\nu(b|s)p(s)ds \tag{7.2}$$

7.2 A 'Least-squares' Encompassing Transition

From section 3.1, a transition probability on $\mathcal{A} \times \mathcal{B}$ can be reinterpreted as a transformation from $[\mathcal{B}]_\infty$ to $[\mathcal{A}]_\infty$. This suggests defining a least-squares encompassing transition $\hat{\Delta}_{\mathcal{B}}^{\mathcal{A}}$ as follows. Consider an ℓ-dimensional random variable $\mathbf{b} \in [\mathcal{B}]$ and a k-dimensional random variable $\mathbf{a} \in [\mathcal{A}]$. Expectations are denoted by the operator E, together with a subscript 1 (2) to denote expectations in $\mathcal{M}_M$ ($\mathcal{N}_N$). The shorthand notation $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ represents $E_1(\mathbf{a}|\mathcal{S})$ and $E_2(\mathbf{b}|\mathcal{S})$ respectively. It is also assumed that $\hat{\mathbf{b}}$ is square-integrable in $\mathcal{M}_M$. As usual in the context of least-squares formulae, all expressions are in deviations from their means.

Definition 7.1. *A least-squares pseudo-true value of* $\mathbf{b}$ *relative to* $\mathbf{a}$ *is a linear expression of the form* $\hat{\mathbf{b}}(\mathbf{a}) = \hat{\mathbf{\Lambda}}'\mathbf{a}$, *where the* $k \times \ell$ *matrix* $\hat{\mathbf{\Lambda}}$ *minimizes:*

$$E_1\left[\left(\mathbf{\Lambda}'\hat{\mathbf{a}} - \hat{\mathbf{b}}\right)'\left(\mathbf{\Lambda}'\hat{\mathbf{a}} - \hat{\mathbf{b}}\right)\right]$$

i.e. is a solution of the linear system:

$$E_1(\hat{\mathbf{a}}\hat{\mathbf{a}}')\hat{\mathbf{\Lambda}} = E_1\left(\hat{\mathbf{a}}\hat{\mathbf{b}}'\right) \tag{7.3}$$

If, in particular, $E_1(\hat{\mathbf{a}}\hat{\mathbf{a}}')$ is non-singular, then $\hat{\mathbf{\Lambda}}$ is unique. Here also the arguments in favour of using least squares pseudo-true values are computational tractability and asymptotic behavior.

Theorem 7.2. *If:*
(i) $\mathcal{M}_M$ is exactly estimable and $E_1(\mathbf{aa}')$ is non-singular;
(ii) $\hat{\mathbf{b}}$ converges Π-a.s. and in $\mathcal{L}^2(\Pi)$ towards $\mathbf{b}(\mathbf{a})$, then:

$$\tilde{\Lambda} \to \Lambda_0 = \left[E_1(\mathbf{aa}')\right]^{-1} E_1\left[\mathbf{ab}'(\mathbf{a})\right] \tag{7.4}$$

Proof: Following (i), $\hat{\mathbf{a}} \to \mathbf{a}, \Pi$-a.s. in $\mathcal{L}^2(\Pi)$. Hence, by Schwartz's inequality $\hat{\mathbf{a}}\hat{\mathbf{a}}' \to \mathbf{aa}'$ and $\hat{\mathbf{a}}\hat{\mathbf{b}}' \to \mathbf{ab}'(\mathbf{a})$ Π-a.s. in $\mathcal{L}^2(\Pi)$ and (7.4) follows. ■

When $\mathbf{b}(\mathbf{a})$ belongs to the set generated by $\mathbf{a}$, i.e. when $\mathbf{b}(\mathbf{a}) = \Lambda_0'\mathbf{a}$, then $\hat{\mathbf{b}}(\mathbf{a}) \to \mathbf{b}(\mathbf{a})$, Π-a.s. in $\mathcal{L}^2(\Pi)$. In general the approximate specificity associated with the least-squares encompassing transition converges towards the norm of the difference between $\mathbf{b}(\mathbf{a})$ and $\Lambda_0'\mathbf{a}$, which is given by:

$$E_1\left[\mathbf{b}(\mathbf{a})\mathbf{b}'(\mathbf{a})\right] - E_1\left[\mathbf{b}(\mathbf{a})\mathbf{a}'\right]\left[E_1(\mathbf{aa}')\right]^{-1} E_1\left[\mathbf{ab}'(\mathbf{a})\right]$$

and will be zero if $\mathbf{b}(\mathbf{a}) = \Lambda_0'\mathbf{a}$.

In general it ought to be possible to select a_is of the form $\mathbb{1}_{A_i}$, in such a way that $\Lambda_0'\mathbf{a}$ is arbitrarily 'close' to $\mathbf{b}(\mathbf{a})$. This essentially follows from our next theorem.

Theorem 7.3. *Let $(\mathcal{A}_\mathcal{K})_{\mathcal{K}\geq 0}$ be a growing sequence of σ-fields such that:*
(i) $\mathcal{A}_\mathcal{K}$ is generated by a partition $(A_1^\mathcal{K} \ldots A_\mathcal{K}^\mathcal{K})$ of A with $\mu_i^\mathcal{K} = \mu_\mathcal{A}(A_i^\mathcal{K}) \neq 0$;
(ii) $\bigvee_{\mathcal{K}\geq 0} \mathcal{A}_\mathcal{K} = \mathcal{A}$.
Let $a_i^\mathcal{K} = \mathbb{1}_{A_i^\mathcal{K}}$. If $b(a) \in \mathcal{L}^2$, then $\hat{b}_\mathcal{K}(a) \to b(a)$ μ-a.s. in $\mathcal{L}^2$

Proof: Under assumptions (i) and (ii), $\hat{b}(a)$ may be rewritten as:

$$\hat{b}_\mathcal{K}(a) = \sum_{i=1}^{\mathcal{K}} \frac{E_1\left[b(a)\,\mathbb{1}_{A_i^\mathcal{K}}\right]}{\mu\left(A_i^\mathcal{K}\right)}\mathbb{1}_{A_i^\mathcal{K}}$$

and the result follows from a martingale theorem (see e.g. Dellacherie and Meyer, 1975). ■

Section 8 applies (7.3) to the choice of regressors problem.

7.3 A 'discrete' encompassing transition

The fact that solutions to (5.4) are, conceptually at least, fairly straightforward when the sample spaces are finite suggests another way of designing approximate encompassing transitions. We can partition A and B into finite numbers of measurable sets, say $(G_i)_{i=1\to m}$ and $(F_j)_{j=1\to n}$ respectively. An approximate discrete encompassing transition then consists of an $m \times n$ matrix $\mathbf{\Delta}$, whose $(i,j)^{th}$ element is a conditional probability for F_j given G_i:

$$\mathbf{\Delta} = \{\delta_{ij}\} \quad \delta_{ij} \geq 0, \ \sum_{j=1}^{m} \delta_{ij} = 1 \tag{7.5}$$

If, for example, we use the Kullback-Leibler criterion, the optimal discrete transition is a solution of the following optimization problem:

$$\min_{\delta_{ij}} E_1 \left\{ \sum_{i,j} \delta_{ij}\mu\left(G_i|S\right) \log \left[\frac{\sum_i \delta_{ij}\mu\left(G_i|S\right)}{\nu\left(F_j|S\right)} \right] \right\} \tag{7.6}$$

subject to the constraints in (7.5). We are presently developing numerical techniques for the evaluation of the expectation (under P_S) in (7.6) and for its solution. Though no formal proofs are offered in the present paper, discrete encompassing transitions should typically have the following properties:

(i) They can be made arbitrarily close to the optimal encompassing transition by refinements of the two partitions;

(ii) Discretisation on A increases the specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$ since it is equivalent to imposing constraints on the form of δ;

(iii) Discretisation on B should instead decrease the specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$ since it 'condenses' the information to be accounted for.

8 An application to the choice of regressors problem

The various concepts discussed in the previous sections are now applied to the choice of regressors problem which was introduced in example 4.3. Only those results that can be derived analytically will be discussed here. See Florens *et al.* [10] for an example of how numerical (simulation) techniques can be used when analytical results are not available. For simplicity, we assume that the variances σ^2 and τ^2 are known.[11] Hence $\mathrm{a} = \beta$ and $\mathrm{b} = \gamma$. The sampling models are those given in (4.6) and (4.7) respectively. The corresponding priors are assumed to be 'Natural Conjugate' priors, as defined e.g. in Raiffa and Schlaifer [30] or Zellner [34]. Hence:

$$\mu_{\mathcal{A}} : \mathrm{a} \sim \mathsf{N}_k\left(\mathrm{a}_0, \sigma^2\mathrm{M}_0^{-1}\right); \quad \nu_{\mathcal{B}} : \mathrm{b} \sim \mathsf{N}_\ell\left(\mathrm{b}_0, \tau^2\mathrm{N}_0^{-1}\right) \tag{8.1}$$

The posterior distributions have similar functional forms with parameters:

$$\mathrm{a}_* = \mathrm{M}_*^{-1}\left(\mathrm{M}_0\mathrm{a}_0 + \mathrm{X}'\mathrm{y}\right) \text{ and } \mathrm{b}_* = \mathrm{N}_*^{-1}\left(\mathrm{N}_0\mathrm{b}_0 + \mathrm{Z}'\mathrm{y}\right) \tag{8.2}$$

$$\mathrm{M}_* = \mathrm{M}_0 + \mathrm{X}'\mathrm{X} \qquad \mathrm{N}_* = \mathrm{N}_0 + \mathrm{Z}'\mathrm{Z} \tag{8.3}$$

[11] Extensions to the case where σ^2 and τ^2 are unknown are currently under study. The transition probabilities $\Delta_{\mathcal{B}}^{\mathcal{A}}$ between the two inferential models are then obtained by reduction of an overall inverted-Wishart density on $\mathcal{A} \vee \mathcal{B}$ in line with our discussion in section 4.2.

The predictive P_S is given by:

$$P_S : \quad \mathbf{y} \sim \mathsf{N}\left(\mathbf{X}\mathbf{a}_0, \sigma^2\left[\mathbf{I}_T + \mathbf{X}\mathbf{M}_0^{-1}\mathbf{X}'\right]\right) \tag{8.4}$$

In the present context, it is natural to define $\mathcal{D}$ as a class of conditional normal distributions of the form:[12]

$$\Delta_{\mathcal{B}}^{\mathcal{A}} : \mathsf{b}|\mathsf{a} \sim \mathsf{N}_\ell(\mathbf{C}\mathbf{a} + \mathbf{c}, \mathbf{V}) \tag{8.5}$$

with $\mathbf{C} \in \mathbb{R}^{\ell \times k}, \mathbf{c} \in \mathbb{R}^\ell$ and $\mathbf{V} \geq 0$.

Let $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ denote the OLS (ML) estimators of $\mathbf{a}$ and $\mathbf{b}$ respectively. The classical (finite sample) pseudo-true value of $\hat{\mathbf{b}}$ on $\mathcal{M}$ is given by $\mathbf{b}(\mathbf{a}) = \hat{\mathbf{\Pi}}\mathbf{a}$, with $\hat{\mathbf{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. The encompassing difference $\hat{\mathbf{b}} - \mathbf{b}(\hat{\mathbf{a}}) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{y}$ is normally distributed on $\mathcal{M}$ with zero mean and covariance matrix $\sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$. As shown e.g. in Mizon and Richard [28], the corresponding Wald encompassing test (WET) statistic is:

$$W(\mathbf{y}) = \frac{1}{\sigma^2}\mathbf{y}'\mathbf{M}_X\mathbf{Z}(\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^+\mathbf{Z}'\mathbf{M}_X\mathbf{y} \underset{\mathcal{M}}{\sim} \chi^2(r) \tag{8.6}$$

where $r = rank(\mathbf{Z}'\mathbf{M}_X\mathbf{Z})$ and $(\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^+$ denotes the Moore-Penrose inverse of $\mathbf{Z}'\mathbf{M}_X\mathbf{Z}$. In the rest of the discussion, we assume - without loss of generality - that $\mathbf{Z}'\mathbf{M}_X\mathbf{Z}$ is non-singular, i.e. that $(\mathbf{X} : \mathbf{Z})$ has rank $k + \ell$.

The $\mathcal{M}$-posterior density of $\mathbf{b}$ is given by:

$$\mu_{\mathcal{A}}^{S} \circ \Delta_{\mathcal{B}}^{\mathcal{A}} \quad : \quad \mathsf{b}|\mathsf{y} \sim \mathsf{N}_\ell(\mathbf{C}\mathbf{a}_* + \mathbf{c}, \mathbf{\Omega}) \tag{8.7}$$

where $\mathbf{\Omega} = \mathbf{V} + \sigma^2\mathbf{C}\mathbf{M}_*^{-1}\mathbf{C}'$. The p-specificity of $\mathcal{N}_N$ is trivial to evaluate for $p = 2$ and for an arbitrary norm matrix $\mathbf{Q} > 0$. It implies the same optimal values for $\mathbf{C}$ and $\mathbf{c}$ as the φ-specificity which is the focus of the following discussion.[13] Following Florens and Scotto [15], the negative entropy between $\nu_{\mathcal{B}}^{S}$ and $\mu_{\mathcal{A}}^{S} \circ \Delta_{\mathcal{B}}^{\mathcal{A}}$ is:

$$\xi_\Delta^E(\mathbf{y}) = \ \tfrac{1}{2}\ \left[\log|\mathbf{\Omega}| - \log\left|\tau^2\mathbf{N}_*^{-1}\right| + \tau^2 tr\left(\mathbf{\Omega}^{-1}\mathbf{N}_*^{-1}\right) + \boldsymbol{\lambda}_*'\mathbf{\Omega}^{-1}\boldsymbol{\lambda}_* - \ell\right] \tag{8.8}$$

with $\boldsymbol{\lambda}_* = \mathbf{b}_* - (\mathbf{C}\mathbf{a}_* + \mathbf{c})$. As discussed in sections 5.4 and 5.5, $\xi_\Delta^E(\mathbf{y})$ can serve as a Bayesian encompassing test statistic. Its expectation on P_S is:

$$\bar{\xi}_\Delta^E = \ \tfrac{1}{2}\ \left[\log|\mathbf{\Omega}| - \log\left|\tau^2\mathbf{N}_*^{-1}\right| + \tau^2 tr\left[\mathbf{\Omega}^{-1}\left(\sigma^2\mathbf{\Phi}_* + \tau^2\mathbf{N}_*^{-1}\right)\right] + \mathbf{m}_*'\mathbf{\Omega}^{-1}\mathbf{m}_* - \ell\right] \tag{8.9}$$

[12] It can be shown that the optimal transition, in the sense of (5.4), has the assumed form within a much broader class of dominated transition probabilities.

[13] That p-specificity does not depend on the transition variance matrix $\mathbf{V}$ which can, therefore, be chosen arbitrarily. To secure desirable asymptotic properties, we should select a sequence $(\mathbf{V})_n$ which tends to zero.

where $\mathbf{m}_*$ and $\sigma^2\mathbf{\Phi}_*$ are the predictive mean and covariance matrix of $\boldsymbol{\lambda}_*$ respectively. Elementary matrix manipulations lead to the following expressions for $\mathbf{m}_*$ and $\mathbf{\Phi}_*$:

$$\mathbf{\Phi}_* = \mathbf{\Sigma}_* + (\mathbf{C} - \mathbf{C}_*)\mathbf{M}_*^{-1}\mathbf{X}'\mathbf{X}\mathbf{M}_*^{-1}(\mathbf{C} - \mathbf{C}_*)' \tag{8.10}$$

$$\mathbf{m}_* = \mathbf{N}_*^{-1}\left(\mathbf{N}_0\mathbf{b}_0 + \mathbf{Z}'\mathbf{X}\mathbf{a}_0\right) - \left(\mathbf{C}\mathbf{a}_0 + \mathbf{c}\right) \tag{8.11}$$

with

$$\mathbf{\Sigma}_* = \mathbf{N}_*^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\mathbf{N}_*^{-1} \tag{8.12}$$

$$\mathbf{C}_* = \mathbf{N}_*^{-1}\hat{\mathbf{Q}}\mathbf{M}_*, \quad \hat{\mathbf{Q}} = \mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{8.13}$$

Hence the values of $\mathbf{C}$ and $\mathbf{c}$ which minimize $\bar{\xi}_\Delta^E$ are given by $\mathbf{C} = \mathbf{C}_*$ and $\mathbf{c} = \mathbf{c}_*$ where:

$$\begin{aligned} \mathbf{c}_* &= \mathbf{N}_*^{-1}\left(\mathbf{N}_0\mathbf{b}_0 + \mathbf{Z}'\mathbf{X}\mathbf{a}_0\right) - \mathbf{C}_*\mathbf{a}_0 \\ &= \mathbf{N}_*^{-1}\left(\mathbf{N}_0\mathbf{b}_0 - \hat{\mathbf{Q}}\mathbf{M}_0\mathbf{a}_0\right) \end{aligned} \tag{8.14}$$

implying $\mathbf{m}_* = \mathbf{0}$ and $\mathbf{\Phi}_* = \mathbf{\Sigma}_*$.

Substituting these values in (8.9) and minimizing with respect to $\mathbf{\Omega}$ yields the following optimal choice for $\mathbf{\Omega}$:

$$\mathbf{\Omega}_* = \sigma^2\mathbf{\Sigma}_* + \tau^2\mathbf{N}_*^{-1} \tag{8.15}$$

whence the φ-specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$ is:

$$\tau_E(\mathcal{N}_N;\mathcal{M}_M) = \tfrac{1}{2}\left[\log|\mathbf{\Omega}_*| - \log\left(\tau^2\left|\mathbf{N}_*^{-1}\right|\right)\right] \tag{8.16}$$

We have implicitly assumed that $\mathbf{V}_* = \mathbf{\Omega}_* - \sigma^2\mathbf{C}_*\mathbf{M}_*^{-1}\mathbf{C}_*' \geq 0$, otherwise the above minimization would have to be subject to the (partially) binding constraint $\mathbf{V} \geq \mathbf{0}$. Note that, in sharp contrast with posterior odds, the φ-specificity in (8.16) is unambiguously defined under non-informative priors ($\mathbf{M}_0 = \mathbf{0}$ and $\mathbf{N}_0 = \mathbf{0}$), in which case:

$$\mathbf{V}_* = (\tau^2 - \sigma^2)\left(\mathbf{Z}'\mathbf{Z}\right)^{-1} + 2\sigma^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \tag{8.17}$$

Hence a sufficient condition for $\mathbf{V}_* > 0$ - at least in 'large sample' situations - is $\tau^2 > \sigma^2$. As discussed in [22], this 'variance dominance' condition plays a key role in the encompassing framework. As the sample size tends to infinity, the optimal transition $\Delta_{\mathcal{B}}^{\mathcal{A}}$ tends to a limiting Dirac probability centered around the classical pseudo-true value $\mathbf{\Pi a}$, where $\mathbf{\Pi} = \operatorname{plim}\hat{\mathbf{\Pi}}$ (under appropriate assumptions on the exogenous process).

Note that the encompassing test statistic $\xi_\Delta^E(\mathbf{y})$, as given in (8.8) is not centered on zero on P_S but, more meaningfully, on the φ-specificity of $\mathcal{N}_N$ relative to $\mathcal{M}_M$. As discussed in Florens and Mouchart [11] in a related context, its distribution on P_S can be expressed in terms of a mixture of $\chi^2(1)$s and standardized normals which could be calibrated by simulation.

The optimal coherent prior $\tilde{\nu}(\mathbf{b})$ which was discussed in section 6.5 cannot be characterized analytically in full. If we substitute the coherent prior mean $\mathbf{C}_*\mathbf{a}_0+\mathbf{c}_*$ for $\mathbf{b}_0$ in (8.14) and solve for $\mathbf{b}_0$, we find that the 'optimal-coherent' prior mean of $\mathbf{b}$ is:

$$\tilde{\mathbf{b}}_0 = \hat{\mathbf{\Pi}}\mathbf{a}_0 \tag{8.18}$$

where $\hat{\mathbf{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. An optimal-coherent prior covariance matrix cannot be obtained analytically. Its numerical derivation goes beyond the objectives of this paper.

The two 'point-optimal' transition probabilities described in section 5.3 are easily obtained. The expectation of $\log \nu(\mathbf{b}|\mathbf{s})$ under the sampling density $p(\mathbf{s}|\mathbf{a})$ takes the form of a normal density as defined in (8.5) with parameters:

$$\tilde{\mathbf{C}} = \mathbf{N}_*^{-1}\mathbf{Z}'\mathbf{X}, \quad \tilde{\mathbf{c}} = \mathbf{N}_*^{-1}\mathbf{N}_0\mathbf{b}_0, \quad \tilde{\mathbf{V}} = \sigma^2\mathbf{N}_*^{-1} \tag{8.19}$$

If $\mathbf{M}_0 = \mathbf{0}$, then $(\tilde{\mathbf{C}}, \tilde{\mathbf{c}}) = (\mathbf{C}_*, \mathbf{c}_*)$. More generally, both pairs share a common large sample limit which is given by $(\hat{\mathbf{\Pi}}, \mathbf{0})$ as in (8.18). A large sample approximation for $\tilde{\mathbf{V}}$ is given by $\sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}$ which differs from the corresponding approximation for $\mathbf{V}_*$ as given in (8.17). The point optimal prior (5.10) associated with the Hellinger distance is also of the form given in (8.5) with parameters $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{c}}$ as defined in (8.19). Its covariance matrix is:

$$\hat{\mathbf{V}} = \sigma^2\mathbf{N}_*^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{N}_*^{-1} + \tau^2\mathbf{N}_*^{-1} \tag{8.20}$$

Finally, we can also obtain analytical expressions for the approximate encompassing transitions that were proposed in sections 7.1 and 7.2. The 'marginalized likelihood' transition, as defined in (7.1), takes the form of the normal distribution in (8.5) with parameters $(\tilde{\mathbf{C}}, \tilde{\mathbf{c}})$ as given in (8.19) and covariance matrix:

$$\hat{\mathbf{V}} = \sigma^2\mathbf{N}_*^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{N}_*^{-1} + \tau^2\mathbf{N}_*^{-1} \tag{8.21}$$

A 'least squares' (LS) pseudo-true value is easily obtained for the random variables $\mathbf{a}_*$ and $\mathbf{b}_*$ in deviations from their sample means on $\mathcal{M}$. Let:

$$\tilde{\mathbf{a}} = \mathbf{M}_*^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{a}), \quad \tilde{\mathbf{b}} = \mathbf{N}_*^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\mathbf{a}) \tag{8.22}$$

whence:

$$E_1(\tilde{\mathbf{a}}\tilde{\mathbf{a}}') = \sigma^2\mathbf{M}_*^{-1}\mathbf{X}'\mathbf{X}\mathbf{M}_*^{-1}, \quad E_1\left(\tilde{\mathbf{a}}\tilde{\mathbf{b}}'\right) = \sigma^2\mathbf{M}_*^{-1}\mathbf{X}'\mathbf{Z}\mathbf{N}_*^{-1} \tag{8.23}$$

It follows that the LS pseudo-true value of $\mathbf{b}_*$ relative to $\mathbf{a}_*$ is given by $\hat{\mathbf{b}}(\mathbf{a}) = \mathbf{C}_*\mathbf{a} + \mathbf{c}_*$, where $\mathbf{C}_*$ and $\mathbf{c}_*$ are given in (8.13) and (8.14) respectively. It is no surprise that in the context of linear regression models, the LS pseudo-true value of $\mathbf{b}_*$ relative to $\mathbf{a}_*$ coincides with the mean of the optimal transition associated with the specificity measure introduced in equations (8.9) to (8.15).

Thus, the various measures of specificity in section 5 and the alternative transitions in section 7 have been shown to be operational in an important model class. While we cannot unequivocally rank these, they converge to the same Dirac transition asymptotically, only differing in their use of prior information.

9 Conclusion

The concept of encompassing (that one model can explain the results obtained by another model) has been considerably developed and formalized since the initial intuitive proposal in Davidson, Hendry, Srba and Yeo [8]. This paper sought to extend its application to the Bayesian approach, and to highlight the resulting close similarities with classical methods. Encompassing was reinterpreted as a concept of sufficiency between inference procedures, 'dual' to that of sufficiency among sampling processes. The main tool was that of a transition probability, which applied to both classical and Bayesian approaches. Exact encompassing supplied the baseline from which departures could be measured by the specificity of the alternative model, and various measures of specificity were introduced. The intractability of the optimal transition probability, namely that which minimized the specificity of the rival model, led to various approximations being considered as operational approaches in practice. These were then applied to the choice of regressors problem to illustrate their differences and commonalties. The relation of encompassing to model choice was discussed.

In summary, encompassing is formalized as a concept of sufficiency among models whereas specificity measures the lack of encompassing. Both concepts are designed to cover classical and Bayesian viewpoints. Tests for encompassing, related to Bayesian posterior odds, are developed. Half a dozen operational transition probabilities are introduced and applied to the choice of regressors problem. They differ in their treatment of prior information relative to $\mathcal{M}_M$ and $\mathcal{N}_N$, but all converge towards the same limiting Dirac distribution, which is centered on the 'classical' pseudo-true value as given in (2.4). A detailed analysis of their relative merits by simulation is currently under investigation. More importantly, however, considerations of analytical and numerical tractability are bound to play a key role outside of the choice of regressors problems. In that respect, the more operational alternatives are likely to be the 'marginalized likelihood' transition (section 7.1), the 'least squares' transition (section 7.2) and, possibly, the 'point optimal transition' (section 5.3).

10 Appendix: Technical details

A.1 Proof of Theorem 4.1

Necessity : Π^* is defined as the "product" of Π and $\Delta_{\mathcal{B}}^{\mathcal{A}}$ in accordance with formula (3.2). Condition (i) follows from that definition. Under $\Pi^*, \Delta_{\mathcal{B}}^{\mathcal{A}}$ also represents the conditional probability on $\mathcal{B}$ given $\mathcal{A} \otimes \mathcal{S}$, whence condition (iii) follows. The marginal probability on $\mathcal{S}$ under Π^* is P_S and $N_{\mathcal{B}}^{*S}$, the correponding transition on $\mathcal{B}$ given $\mathcal{S}$, is then defined by the following identity:

$$\forall F \in \mathcal{B},\ X \in \mathcal{S},\ \Pi^*(A \times F \times X) = \int_X N_{\mathcal{B}}^{*S}(F) dP_S \qquad (A.1)$$

On the other hand, if follows from the definition of Π^* that:

$$\begin{aligned}\forall F \in \mathcal{B},\ X \in \mathcal{S},\ \Pi^*(A \times F \times X) &= \int_X \left[\int_A \Delta_{\mathcal{B}}^{\mathcal{A}}(F) dM_{\mathcal{A}}^{S}\right] dP_S \\ &= \int_X (M_{\mathcal{A}}^{\mathcal{S}} o \Delta_{\mathcal{S}}^{\mathcal{A}})(F) dP_S\end{aligned} \qquad (A.2)$$

Condition (ii) follows by comparison between formula (A.1) and (A.2).

Sufficiency : Π^* being given, a version of the conditional probability on $\mathcal{B}$ given $\mathcal{A} \otimes \mathcal{S}$ which does not depend on $\mathcal{S}$ provides a transition $\Delta_{\mathcal{B}}^{\mathcal{A}}$ which satifies definition 4.1 in accordance with the decompositions in formulae (A.1) and (A.2). ■

A.2 Proof of Theorem 4.2

Necessity : The probability Ψ is constructed as follows:

- The marginal of Ψ on $\mathcal{I}$ is defined by $\alpha(1)$ and $\alpha(2) > 0$;

- Conditionally on $i = 1$, the distribution on $(\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}$ is taken to be Π^*, as defined in theorem 4.1;

- Conditionally on $i = 2$, a distribution χ^* on $(\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}$ is defined as follows:

$$\forall E \in \mathcal{A}, F \in \mathcal{B}, X \in \mathcal{S}, \chi^*(E \times F \times X) = \int_{F \times X} \Omega_{\mathcal{A}}^{\mathcal{B}}(E) d\chi \qquad (A.3)$$

where $\Omega_{\mathcal{A}}^{\mathcal{B}}$ is an arbitrary transition on $\mathcal{B} \times \mathcal{A}$.

Conditions (i) to (iii) are verified by construction. Condition (iv) follows from the fact that under Π^* $(\chi^*), \Delta_{\mathcal{B}}^{\mathcal{A}}$ $(\Omega_{\mathcal{A}}^{\mathcal{B}})$ is a version of a conditional probability on $\mathcal{B}$ (A) conditionally on $\mathcal{A} \otimes \mathcal{S}$ $(\mathcal{B} \otimes \mathcal{S})$. Finally the transitions on B given $\mathcal{S}$ and $\mathcal{I}$ are $M_{\mathcal{A}}^{\mathcal{S}} o \Delta_{\mathcal{B}}^{\mathcal{A}}$ for $i = 1$ and $N_{\mathcal{B}}^{\mathcal{S}} = N_{\mathcal{B}}^{*S}$ for $i = 2$. Hence condition (v) follows from definition 4.1.

A.3 Proof of Theorem 4.3

The notation and the definition of a probability Π^* on $(\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}$ are those introduced in theorem 4.1. Under formula (A.1) the encompassing condition $N_{\mathcal{B}}^{\mathcal{S}} = N_{\mathcal{B}}^{*\mathcal{S}}$ is equivalent to the following condition:

$$\frac{d\Pi^*_{\mathcal{B}\otimes\mathcal{S}}}{d\chi} = \frac{dP_{\mathcal{S}}}{dQ_{\mathcal{S}}} \tag{A.4}$$

where $\Pi^*_{\mathcal{B}\otimes\mathcal{S}}$ denotes the restriction of Π^* on $(\mathcal{B} \otimes \mathcal{S})$. Equivalent reformulations of (A.4) are the following:

$$\Pi^*(A \times F \times X) = \int_{F\times X} \frac{dP_{\mathcal{S}}}{dQ_{\mathcal{S}}} d\chi, \quad \text{and}$$

$$\int_F \left[\int_A P_{\mathcal{S}}^{\mathcal{A}}(X) dK_{\mathcal{A}}^{\mathcal{B}}\right] d\rho_{\mathcal{B}} = \int_F \left[\int_X \frac{dP_{\mathcal{S}}}{dQ_{\mathcal{S}}}\right] d\nu_{\mathcal{B}}$$

$$= \int_F \left[\frac{d\nu_{\mathcal{B}}}{d\rho_{\mathcal{B}}} \int_X \frac{dP_{\S}}{dQ_{\mathcal{S}}} dQ_{\mathcal{S}}^{\mathcal{B}}\right] d\varrho_{\mathcal{B}}$$

wherefrom formula (4.12) and theorem 4.3 follow. ■

A.4 Proof of Theorem 4.5

Necessity : Let Π^* denote the distribution on $(\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}$ introduced in theorem 4.1. The restriction of Π^* on $\mathcal{B} \otimes \mathcal{S}$ equals χ since, under definition 4.5 both imply a common conditional distribution on $\mathcal{B}$ given $\mathcal{S}$ and a common marginal distribution on $\mathcal{S}$. Let $K_{\mathcal{A}}^{\mathcal{B}}$ denote the conditional distribution on $\mathcal{A}$ given $\mathcal{B}$ which is derived from Π^*. Note that $P_{\mathcal{S}}^{\mathcal{A}} \equiv P_{\mathcal{S}}^{\mathcal{A}\otimes\mathcal{B}}$. Condition (i) and (ii) follow.

Sufficiency : We define a probability Γ on $\mathcal{A} \times \mathcal{B}$ as follows:

$$\Gamma(E \times F) = \int_F K_{\mathcal{A}}^{\mathcal{B}}(E) d\nu_{\mathcal{B}}$$

The marginal of Γ on A is $\mu_{\mathcal{A}}$ by condition (ii). From Γ we derive a conditional distribution on $\mathcal{B}$ given $\mathcal{A}$, which is denoted $\Delta_{\mathcal{B}}^{\mathcal{A}}$, and use it next to construct Π^* on $(\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}$ as in theorem 4.1. The proof is completed by establishing that the restriction of Π^* on $\mathcal{B} \otimes \mathcal{S}$ equals χ. We have successively:

$$\chi(F \times X) = \int_F Q_{\mathcal{S}}^{\mathcal{B}}(X) d\nu_{\mathcal{B}} = \int_F \left[\int_A P_{\mathcal{S}}^{\mathcal{A}}(X) dK'^{\mathcal{B}}_{\mathcal{A}}\right] d\nu_{\mathcal{B}}$$

$$= \int_{A\times F} P_{\mathcal{S}}^{\mathcal{A}}(X) d\Gamma = \int_A \Delta_{\mathcal{B}}^{\mathcal{A}}(F) P_{\mathcal{S}}^{\mathcal{A}}(X) d\mu_{\mathcal{A}}$$

$$= \int_{A\times F} \Delta_{\mathcal{B}}^{\mathcal{A}}(F) d\Pi = \Pi^*(A \times F \times S) \quad ■$$

References

[1] Berger, J. (1990), 'Robust Bayesian Analysis: Sensitivity to the Prior', *The Journal of Statistical Planning and Inference*, pp. 303-328.

[2] Berk, R.H. (1966), 'Limiting Behavior of Posterior Distributions when the Model is Incorrect', *The Annals of Mathematical Statistics*, **37**, pp. 51-58.

[3] Berk, R.H. (1970), 'Consistency a Posteriori', *The Annals of Mathematical Statistics*, **41**, pp. 894-906.

[4] Billingsley, P. (1968), *Convergence of Probability Measures.* New York: John Wiley and Sons.

[5] Blackwell, D. (1951), 'Comparison of Experiments', *Proceedings of the second Berkeley Symposium of Mathematical Statistics and Probability*, University of California Press, pp. 93-102.

[6] Blackwell, D. (1953), 'Equivalent Comparison of Experiments', *Annals of Mathematical Statistics*, **24**, pp. 265-272.

[7] Czisar, I. (1967), 'On Information Type Measures of Difference of Probability Distributions and Indirect Observations'. *Studia Scientiarum Mathematicarum Hungria*, **2**, pp. 299-318.

[8] Davidson, J.E.H., Hendry, D.F., Srba, F. and Yeo, S. (1978), 'Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom', *Economic Journal*, **88**, 661-92.

[9] Dellacherie C. and P.A. Meyer (1975). *Probabilité et Potentiel*, Paris: Hermann. English translation: *Probabilities and Potential* (1978). New York: North-Holland.

[10] Florens, J.P., S. Larribeau and M. Mouchart (1992), 'Bayesian Encompassing Test of a Unit Root Hypothesis', Cahier du Gremaq 9227, Universit de Toulouse (France).

[11] Florens, J.P. and M. Mouchart (1989), 'Bayesian Specification Tests', in B. Cornet and H. Tulkens (eds.) *Contributions to Operations Research and Econometrics.* Cambridge (Mass): MIT Press.

[12] Florens, J.P. and M. Mouchart (1993), 'Bayesian Testing and Testing Bayesians', in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics vol 11: Econometrics*, Amsterdam: North Holland.

[13] Florens, J.P., Mouchart, M. and Rolin, J-M . (1990), *Elements of Bayesian Statistics.* New York: Marcel Dekker.

[14] Florens, J.P. and J.F. Richard (1989), 'Encompassing in Finite Parametric Spaces', Mimeo, Duke Unversity.

[15] Florens, J.P. and S. Scotto (1984), 'Information Value and Econometric Modelling', *Southern European Economic Discussion Paper Series*, **17**, GREQE,

University of Aix-Marseille.

[16] Goel, P.K. and M.H. DeGroot (1979), 'Comparison of Experiments and Information Measures', *Annals of Statistics*, **7**, 5, pp. 1066-1077.

[17] Gouriéroux, C., A. Monfort and A. Trognon (1983), 'Testing Nested or Nonnested Hypotheses'. *Journal of Econometrics*, **38**, pp. 73-90.

[18] Gouriéroux, C., Monfort, A. and Trognon, A. (1984), 'Pseudo-Maximum Likelihood Methods: Theory', *Econometrica*, **52**, 681-700.

[19] Govaerts, B., Hendry, D.F. and Richard, J-F. (1993), 'Encompassing in Stationary Linear Dynamic Models', forthcoming, *Journal of Econometrics.*

[20] Hendry, D.F. and Richard, J.F. (1982), 'On the Formulation of Empirical Models in Dynamic Econometrics', *Journal of Econometrics*, **20**, 3-33. Reprinted as p304-334 in C.W.J. Granger (ed.) (1990), *Modelling Economic Series*, Oxford: Clarendon Press.

[21] Hendry, D.F. and Richard, J.F. (1983), 'The Econometric Analysis of Economic Time Series' (with discussion), *International Statistical Review*, **51**, 111-63.

[22] Hendry, D.F. and Richard, J.F. (1989), 'Recent Developments in the Theory of Encompassing', p393-440 in B. Cornet and H. Tulkens (eds.), *Contributions to Operations Research and Econometrics. The XXth Anniversary of CORE.* Cambridge (Mass): MIT Press.

[23] Huber, P.J. (1967), 'The Behavior of Maximum Likelihood Estimates under Non-Standard Conditions', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, Berkeley, The University of California Press.

[24] Kiefer, N. and J.F. Richard (1987), 'Decision Theory, Estimation Strategies and Model Choice', CAE Working Paper 87-08. Cornell University.

[25] Lavine, M. (1991), 'Sensitivity and Bayesian Statistics: the Prior and the Likelihood', *Journal of the American Statistical Association*, **86**, pp. 396-399.

[26] Lecam, L. (1964), 'Sufficiency and Approximate Sufficiency', *Annals of Mathematical Statistics*, **35**, pp. 1419-1455.

[27] Mizon, G.E. (1984), 'The Encompassing Approach in Econometrics', p135-72 in D.F. Hendry and K.F. Wallis (eds.), *Econometrics and Quantitative Economics.* Oxford: Basil Blackwell.

[28] Mizon G.E. and Richard, J-F. (1986), 'The Encompassing Principle and its Application to Non-Nested Hypothesis Tests', *Econometrica*, **54**, 657-78.

[29] Neveu, J. (1970). *Bases Mathmatiques des Probabilits*, Paris: Masson (2nd edition). English translation: *Mathematical Foundations of the Calculus of Probability* (1965). San Francisco: Holden-Day.

[30] Raiffa, H. and R. Schlaifer (1961), *Applied Statistical Decision Theory.* Boston (Mass): Division of Research, Harvard Business School.

[31] Sawa, T. (1978), 'Information Criteria for Discriminating Among Alternative Regression Models', *Econometrica*, **46**, 1273-92.

[32] Torgensen, E.N. (1976), 'Comparison of Statistical Experiments', *Scandinavian Journal of Statistics*, **3**, pp. 186-208.

[33] White, H. (1982), 'Maximum likelihood Estimation of Misspecified Models', *Econometrica*, **50**, 1-26.

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