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 $S = \left\{\mathbf{y}_i \in I\!R^2; i = 1, \ldots, n\right\}$ 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:

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

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 $\mathbf{H} = \mathbf{\Sigma}^{-1} = (h_{ij})$. 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:

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

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:

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

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 $\tilde{\mathcal{M}}$. 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]

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

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

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

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

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

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

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

$$\forall Z \in \mathcal{W},\quad \Gamma_{\mathcal{W}}^{\mathcal{U}}(u,Z) = \int_V \Delta_{\mathcal{W}}^{\mathcal{V}}(v,Z) \Lambda_{\mathcal{V}}^{\mathcal{U}}(u,dv) \tag{3.5}$$

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):

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

[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 $\Lambda_{\mathcal{V}}^{\mathcal{U}}$ 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 $\lambda : (U,\mathcal{U}) \to (V,\mathcal{V})$ 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 $P^{\mathcal{A}}_{\mathcal{S}}$.*

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	$\mathcal{M} = \{(A,\mathcal{A}),(S,\mathcal{S}),P^{\mathcal{A}}_{\mathcal{S}}\}$	$\mathcal{N} = \{(B,\mathcal{B}),(T,\mathcal{T}),Q^{\mathcal{B}}_{\mathcal{T}}\}$
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_S^A and Q_T^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_A^S and N_B^S. Examples of estimation procedures are:

(a) Estimators: if $\hat{a} : (S, \mathcal{S}) \to (A, \mathcal{A})$ is an 'estimator', then the Dirac measure $D_{A,\hat{a}}^S$ 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 μ_A, then the corresponding posterior density μ_A^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 Λ_T^S such that:*

$$Q_T^A = P_S^A \circ \Lambda_T^S \tag{3.14}$$

If, in a Bayesian framework, a common prior probability μ_A is associated with the two sampling models, then the sufficiency condition has to hold μ_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 $P_{S\otimes T}^A$ is an 'extension' of P_S^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) $P_{S\otimes T}^A | \mathcal{T}$ restricted to $\mathcal{T}$ equals Q_T^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_S^0 on $(S, \mathcal{S})$. The choice of P_S^0, which essentially serves to characterize the relevant null sets, often depends on the context. Natural choices are P_S^A from a classical viewpoint or P_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_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_S^0) if and only if there exists a transition probability Δ_B^A, called the pseudo-true value of $\mathcal{N}_N$ within $\mathcal{M}_M$, such that:*

$$N_B^S = M_A^S \otimes \Delta_B^A, \qquad P_S^0 \text{ - a.s.} \tag{4.1}$$

Lemma 4.1. *If $\mathcal{M}_M$ exactly encompasses $\mathcal{N}_N$ (on P_S^0) with pseudo-true value Δ_B^A, if $\mathcal{N}_N$ exactly encompasses $\mathcal{O}_O = (\mathcal{O}, O_C^S)$ (on Q_S^0) with pseudo-true value Λ_C^B and if P_S^0 and Q_S^0 are equivalent, then $\mathcal{M}_M$ exactly encompasses $\mathcal{O}_O$ (on P_S^0) with pseudo-true value:*

$$\Gamma_C^A = \Delta_B^A \circ \Lambda_C^B, \qquad P_S^0 \text{ - a.s.} \tag{4.2}$$

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_A^S and N_B^S are Dirac measures associated with a pair of estimators, $\tilde{a}$ and $\tilde{b}$ respectively. An *additional* restriction is imposed, namely that Δ_B^A is itself a Dirac measure.

Under these conditions, (4.1) simplifies to:

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

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

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

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 $\boldsymbol{\Omega}_{\tilde{b}} = \mathbf{C}\boldsymbol{\Sigma}_{\tilde{a}}\mathbf{C}' + \mathbf{V}$. In such a case, $\mathbf{V}$ measures the loss of efficiency when $\mathbf{Ca} + \mathbf{c}$ 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 $P_{\mathcal{S}}^0$) if and only if there exists a transition probability $\Delta_{\mathcal{B}_1}^{\mathcal{A}\otimes\mathcal{S}_1}$ such that:*

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

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

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

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

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 $\gamma_\beta = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}$. 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_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:*

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

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_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:*

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

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

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

■

Condition (4.9) is often associated with a factorization of the sampling probability $P_S^{\mathcal{A}}$ into a marginal probability $P_{S_1}^{\mathcal{A}}$ and a conditional one $P_S^{\mathcal{A}\otimes\mathcal{S}_1}$ in such a way that:

$$P_{S_1}^{\mathcal{A}} = P_{S_1}^{\mathcal{A}_1} \quad \text{and} \quad P_S^{\mathcal{A}\otimes\mathcal{S}_1} = P_S^{\mathcal{A}_2\otimes\mathcal{S}_1} \tag{4.11}$$

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 Δ_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) $N_B^{\mathcal{S}} = N_B^{*\mathcal{S}}$, where $N_B^{*\mathcal{S}}$ 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 $\{I \times \Theta \times S, \mathcal{I} \otimes (\mathcal{A} \vee \mathcal{B}) \otimes \mathcal{S}\}$. 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. *$\mathcal{M}_M^\mu$ exactly encompasses $\mathcal{N}_N^\nu$ 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) $\mathcal{B} \perp\!\!\!\perp \mathcal{I} | S$ 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 $\mathcal{M}_M^\mu$ and $\mathcal{N}_N^\nu$ 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 $\mathcal{M}_M^\mu$ exactly encompasses $\mathcal{N}_N^\nu$ 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. *$\mathcal{M}_M^\mu$ coherently (exactly) encompasses $\mathcal{N}_N^\nu$ if and only if:*
(i) $\mathcal{M}_M^\nu$ exactly encompasses $\mathcal{N}_N^\nu$ 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:*

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

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

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.