ENCOMPASSING IN FINITE PARAMETRIC SPACES

Jean-Pierre Florens
GREMAQ, Toulouse University
Jean-Francois Richard
ISDS, Duke University
DP #89-03

Abstract

The Bayesian notion of encompassing essentially represents a concept of sufficiency among models. Analytical results can only be derived for a limited class of models. In the present note we analyse the case where the parametric spaces associated with the competing models are finite. This restriction greatly simplifies the analysis and, in particular, enables us to study the asymptotic properties of encompassing transitions by means of elementary probability tools.

Acknowledgement: The second author is supported by a grant of the National Science Foundation (SES-8708615). We are indebted to D.F. Hendry, M. Lavine, M. Mouchart and J.M. Rolin for numerous helpful discussions. All errors are our responsibility.
1. Introduction

In a loose sense, a model $M_1$ encompasses a "rival" model $M_2$ if the former can account for the results obtained by the latter. In the expanding literature on Model Choice, this notion of encompassing has proved to be a powerful tool of analysis. See Hendry and Richard (1989) for a recent survey, as well as for a list of references, or Florens et al. (1989) for a bayesian presentation which is directly relevant for the object of the present note.

The conceptual strength of this bayesian notion of encompassing lies in its interpretation as sufficiency on the parameter space.\(^1\) Specific techniques of application build upon a long sequence of statistical papers relative to the properties of "estimation procedures" for a "mis-specified" model (such as $M_2$ is from the point of view of $M_1$). See, in particular, Cox (1961, 1962), Huber (1967) or, for an asymptotic bayesian presentation, Berk (1966, 1970) who builds upon the earlier contribution of Lecam (1953). See also in the recent econometric literature White (1982), Gourieroux et al. (1983, 1984) or Holly (1987).

Unfortunately, the practical implementation of an encompassing analysis can be tedious and the scope for analytical result is limited. Work is in progress regarding the design of operational numerical encompassing procedures, e.g. in the spirit of the Monte-Carlo sampling procedure outlined in Hendry and Richard (1989, section 5). However, relying upon numerical procedures does not contribute much to bettering our understanding of the more fundamental issues on hand.

Hence, in the present note we propose instead to restrict our attention to the simpler case where the parameter spaces associated with the two rival models are finite. This limitation enables us to investigate analytically the concept of encompassing and, in particular, its asymptotic properties by means of elementary probability tools. Such analysis usefully contributes clarifying the conceptual problems on hand and, thereby, provides guidelines for the largely numerical analysis of more complex situations.

For the ease of presentation we shall assume further that the observations are identically independently distributed (i.i.d.) under either $M_1$ or $M_2$. The latter restriction is not essential and could be replaced by the more fundamental assumption that $M_1$ satisfies whatever stationarity and ergodicity assumptions are required to validate the use of the strong law of large numbers. $M_2$ could be Markovian, or autoregressive of order larger than one. Explanatory (conditioning) variables could be introduced in either model under appropriate assumptions relative to their asymptotic behavior. See e.g. Burgue et al. (1982).

In order to be more explicit on the object of the present note let us briefly summarize the bayesian notion of encompassing, as discussed in greater details by Florens et al. (1989). The two bayesian models $M_1$ and $M_2$ share a common sampling space $S_T$ where $T$ is the sample size, with respective parameter spaces $A$ and $B$. Hence, they are characterized by (joint) probability measures, say $\Pi$ and $\chi$ on the product spaces $A \times S_T$ and $B \times S_T$ respectively. The key step in our encompassing analysis consists in extending $\Pi$ into a probability $\Pi^*$ on the space $A \times B \times S_T$ in such a way that the following two conditions are satisfied:

(i) The parameter of $M_1$ remains sufficient under $\Pi^*$, i.e. the sampling probability derived from $\Pi^*$, conditionally on the elements of $A$ and $B$, depends solely on $A$. Equivalently, $\Pi^*$ is defined as the product of $\Pi$ and of a transition probability $\Delta_T$ on $B$ conditionally on $a \epsilon A$ (independently of $s_T \epsilon S_T$).

---

\(^1\) Sufficiency on the parameter space being itself dual to the classic notion of sufficiency on the sampling space, as introduced by Blackwell (1951, 1953) and extensively analyzed by Lecam (1964, 1986). The concept of duality to which we refer is that whereby the respective roles of parameters and observations are interchanged in the definitions.
\( \Delta_T \) is called a “bayesian pseudo-true value”;

(ii) The posterior probability on \( B \) derived from \( \Pi^* \) is as “close” as possible to that associated with \( \chi \). Specifically we shall select whichever pseudo-true values \( \Delta_T^* \) minimizes an appropriate measure of the divergence between the two posterior probabilities on \( B \) (precise definitions are provided below). If these two probabilities coincide under \( \Delta_T^* \), then \( M_1 \) is said to encompass (exactly) \( M_2 \). Otherwise, the minimal divergence is called the specificity of \( M_2 \) relative to \( M_1 \) and can serve as the basis of a (bayesian) test of the “validity” of \( M_1 \).

In the present paper, we shall discuss the following issues (in the context of finite parameter spaces):

(i) How can we effectively construct bayesian pseudo-true values for a given sample size?

(ii) What are the limits of these pseudo-true values as \( T \) tends to infinity?

(iii) How are these limits related to the limiting posterior probability on \( B \), as derived from the “mis-specified” model \( M_2 \) when \( M_1 \) is assumed to be correctly specified?

The paper is organised as follows: The posterior distributions for the two models are derived in section 2, where we also discuss their limiting behavior under \( M_1 \). Exact and approximate finite sample (Bayesian) pseudo-true value are evaluated in section 3. In section 4 we demonstrate that the specificity of \( M_2 \) relative to \( M_1 \) tends to zero under \( M_1 \) as the sample size increases. Proofs are regrouped in a technical appendix.

2. Posterior Distributions

2.1 Notation and assumptions

The notations for the two models \( M_1 \) and \( M_2 \) are collected together in table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>( M_1 )</th>
<th>( M_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>( \mathcal{I}A = {i, \ldots, m} )</td>
<td>( \mathcal{I}B = {i, \ldots, n} )</td>
</tr>
<tr>
<td>Sample</td>
<td>( s_T = (x_1, \ldots, x_T); x_i \in \mathcal{X}, \forall t; s_T \in \mathcal{S}_T = \mathcal{X}^T )</td>
<td></td>
</tr>
<tr>
<td>Prior probability</td>
<td>( \mu = (\mu(1), \ldots, \mu(m))' ) ( 1 \geq \mu(i) &gt; 0, \Sigma_i \mu(i) = 1 )</td>
<td>( \nu = (\nu(1), \ldots, \nu(n))' ) ( 1 \geq \nu(j) &gt; 0, \Sigma_j \nu(j) = 1 )</td>
</tr>
<tr>
<td>Observation density</td>
<td>( f(x_i</td>
<td>i) )</td>
</tr>
<tr>
<td>Sample density Sample probability</td>
<td>( p(s_T</td>
<td>i) = \Pi_i f(s_T</td>
</tr>
<tr>
<td>Predictive density Predictive probability</td>
<td>( p(s_T) = \Sigma_i p(s_T</td>
<td>i) \mu(i) ) ( P )</td>
</tr>
<tr>
<td>Posterior probability</td>
<td>( \mu_T = (\mu(1</td>
<td>s_T), \ldots, \mu(m</td>
</tr>
<tr>
<td>Joint density Joint probability</td>
<td>( \Pi(s_T,i) = p(s_T</td>
<td>i) \mu(i) )</td>
</tr>
</tbody>
</table>

Table 1: Notation

Table 1 incorporates a number of implicit assumptions and calls for the following additional comments:

(i) The assumption that \( \mu(i) > 0, \forall i \) contributes simplifying our presentation of asymptotic results. In practice, it is hardly restrictive and implies that only those values of the parameter \( i \) which have non zero prior probabilities are retained. The possibility that \( \nu(j) \) can be zero raises no problems at all;
(ii) We shall specifically discuss the identification status of model 1 in the course of our analysis and we shall, in particular, analyze the implications of a lack of identification characterized by the existence of pairs \((i, i')\), with \(i \neq i'\) for which \(f(x_i | i) = f(x_i | i')\) a.s. The identification status of \(M_2\) is inconsequential since the properties of \(M_2\) will be discussed within the context of \(M_1\).

(iii) We shall assume that the predictive density of \(M_2\) dominates that of \(M_1\). In doing so we can discuss properties of \(\nu_T\) that are almost sure relative to \(P\), within the context of \(M_1\), while initially \(\nu_T\) is only defined almost surely relative to \(Q\).

(iv) Since we are dealing with finite parameter spaces, there is no loss of generality in representing the probabilities on \(X\) and \(S_T\) by their respective densities relative to a common measure whose differential components are denoted by \(dx_i\) and \(ds_T\).

2.2 Limiting behavior of the posterior densities \(\mu_T\) and \(\nu_T\)

Following e.g. Huber (1967) or Lecam (1953, 1980), a classical reinterpretation of the parameter \(j\) within the context of \(M_1\) is obtained by minimizing the Kullback-Leibler Information Criteria (KLIC). See Kullback (1959). More specifically, let \(\xi(j|i)\) denote the KLIC of \(q(s_T|j)\) relative to \(p(s_T|i)\):

\[
\xi(j|i) = \int_X \log \left[ \frac{f(x_i | j)}{q(x_i | j)} \right] f(x_i | j) dx_i
\]

\[
= \frac{1}{T} \int_{S_T} \log \left[ \frac{p(s_T | j)}{q(s_T | j)} \right] p(s_T | j) ds_T \tag{2.1}
\]

To every \(i \in A\), we associate a subset of \(B\), consisting of all the \(j\)'s that minimize \(\xi(j|i)\) for the given \(i\):

\[B_i = \{ j \in B \mid \forall j' \in B, \quad \xi(j' | i) \geq \xi(j | i) \}\]

Next, we introduce an \(m \times n\) matrix \(D = (d_{ij})\) whose elements are defined as follows:

\[d_{ij} = \nu(j | B_i) = \begin{cases} \nu(j) / \nu(B_i), & \text{if } j \in B_i \\ 0, & \text{otherwise} \end{cases} \tag{2.3}\]

Note that \(D\) is a transition matrix in the sense that each row of \(D\) is a probability on \(B\), conditionally on the \(B_i\) associated with the row index (\(D\) is a Markovian matrix). In the rest of our discussion we refer to \(D\) as to the (classical) pseudo-true value,\(^1\) following thereby the terminology of Sawa (1978) - see also Gouriou et al. (1984).

The fact that \(B_i\) need not be a singleton deserves careful consideration. It can result from two distinct causes which have fundamentally different implications. Either there might exist several sampling probabilities in \(M_2\) that are equidistant under the KLIC from a given sampling probability in \(M_1\), or there exist a single such probability, but a lack of identification in \(M_2\) associates that probability with several values of the parameter \(j\). As discussed in Berk (1966, 1970), the first possibility has important implications regarding the limiting behavior of \(\nu_T\). The following example usefully illustrates the issue on hand:

Example: Let

\[M_1: A = \{1, 3\} \text{ and } x_i | i \sim N(i, 1) \tag{2.4}\]

\[M_2: B = \{2, 4, 6\} \text{ and } x_i | j \sim N(j, 1) \tag{2.5}\]

\(^1\) Definition (2.3) coincides with the classical definition of a pseudo-true value when the \(B_i\)'s are singletons. It is, however, more general than the latter since it also applies to cases where some \(B_i\)'s contains more than one element. Bayesians will naturally weight distinct \(j\)'s within a common \(B_i\) according to the (conditional) prior probabilities \(\nu(j | B_i)\). The discussion in section 4 below provides additional flesh to definition (2.3).
Hence $\xi(j|i) = \frac{1}{2}(i - j)^2$ and $B_3 = \{2, 4\}$. Also

$$\nu(2|\pi) = \left(1 + \frac{\nu(4)}{\nu(2)} \exp[2T(\pi - 3)] + \frac{\nu(6)}{\nu(2)} \exp[4T(\pi - 4)]\right)^{-1} \tag{2.6}$$

If the true model is characterized by $i = 3$, then the last term in the right-hand of formula (2.6) tends to zero. The second term, however, does not converge since $T(\pi - 3)$ is a centered random walk such that $\liminf T(\pi - 3) = -\infty$ and $\limsup T(\pi - 3) = +\infty$, see Stout (1974, Chp. 6). It follows that, under $M_1$ with $i = 3$, $\liminf \nu(2|\pi) = 0$ and $\limsup \nu(2|\pi) = 1$.

As stated in theorem 1 below, we can establish that $\nu(B_i|\pi)$ tends to 1, $P_i$ almost surely, though, as the example illustrates, $\nu(j|\pi)$ can widely fluctuate for $j$'s indexing distinct probabilities within a single $B_i$. Whenever we wish to eliminate the latter possibility, largely for the ease of presentation, we shall introduce the additional assumption $H$ that the probability which minimizes $\xi(j|i)$ in formula (2.1) is unique for any given $i$:

$$H : \forall i \in A \quad j, j' \in B_i \Rightarrow g(x_i|j) = g(x_i|j') \quad a.s. \tag{2.7}$$

$H$ does not imply that $B_i$ is a singleton since there might still be several $j$'s indexing the unique probability associated with $B_i$ due to a lack of identification of $M_2$.

This being said we can now establish a fundamental theorem relative to the asymptotic behavior of $\mu_\pi$ and $\nu_\pi$ when the "true" sampling probability is $P_i$. Since we wish to account for the fact that $i$ might not be identified, let $A_i$ denote the equivalence class of $i$ for the identification relationship on $M_1$.

**Theorem 1**: Under the assumptions of section 2.1:

(i) $\forall i \in A, \mu(i'|\pi) \rightarrow \mu(i'|A_i), P_i a.s.$

(ii.a) $\forall j \notin B_i, \nu(j|\pi) \rightarrow 0, P_i a.s.$ or equivalently

(ii.b) $\nu(B_i|\pi) \rightarrow 1, P_i a.s.$

If, in addition, assumption $H$ in (2.7) applies, then

(iii) $\forall j \in B_i, \nu(j|\pi) \rightarrow d_{ij}, P_i a.s.$

**Proof**: See Appendix.

3. Finite Sample Pseudo-true Values

In line with our discussion in section 1, an encompassing transition probability or (bayesian) pseudo-true value is defined as a transition probability $\Delta = (\delta_{ij})$, with $i : 1 \rightarrow m$ and $j : 1 \rightarrow n$, such that for each $i$ the vector $(\delta_{i1}, \ldots, \delta_{in})$ is a probability on $B$, conditionally on $i$ ($\delta_{ij} \geq 0$ and $\Sigma_j \delta_{ij} = 1$). Following our assumption that $i$ remains a sufficient parameter under the probability $\Pi^*$ on $A \times B \times S_\pi$, the latter is necessarily given by the product of $\Pi$ and $\Delta$. Under $\Pi^*$, the sampling probability remains $p(s_\pi|i)$ and the prior probabilities on $A \times B$ are given by

$$\mu^*(i, j) = \mu(i) \delta_{ij} \tag{3.1}$$

It follows that, under $\Pi^*$, the posterior probabilities of $j$, conditionally and unconditionally on $i$, are respectively given by

$$\nu^*(j|i, \pi) = \delta_{ij} \tag{3.2}$$

$$\nu^*(j|\pi) = \Sigma_i \delta_{ij} \mu^*(i|\pi) \tag{3.3}$$

Let $\nu^*_\pi$ denote the vector $(\nu^*_1(\pi), \ldots, \nu^*_n(\pi))'$ of posterior probabilities for $j$ under $\Pi^*$. Formula (3.3) is then rewritten as

$$\nu^*_\pi = \Delta' \mu^* \tag{3.4}$$
In summary, $\Pi^*$ represents an extension of $\Pi$ whereby $j$ is incorporated into the parameter space of $M_1$ and, at the same time, the assertion that $M_1$ appropriately characterizes the sampling process is maintained.

We next discuss the choice of $\Delta$ and hence, of $\Pi^*$. Within the encompassing framework, whose object is the analysis of whether or not $M_1$ is "sufficient" relative to $M_2$, we shall aim at selecting a $\Delta$ such that the corresponding $\nu_\Delta^*$ is as "close" as possible to $\nu_T$. If there exists a $\Delta$ such that $\nu_\Delta^*$ and $\nu_T$ are identical to each other for almost all $s_T$'s, then $M_1$ is said to exactly encompass $M_2$. In general, this will not be the case. We may then select a "distance" criterion between $\nu_\Delta^*$ and $\nu_T$, which naturally depends on $s_T$ and $\Delta$ and, hence is denoted $\Psi(\Delta; s_T)$. The chosen $\Delta$ will be one which minimizes the expectation of $\Psi$ under the predictive probability of $P$ of $M_1$:

$$\Delta_0^* = \arg \max_{\Delta} \mathbb{E}_P[\Psi(\Delta; s_T)]$$

The minimization is subject to the constraint that $\Delta$ is a transition probability, i.e. that $\delta_{ij} \geq 0$ and $\Sigma_i \delta_{ij} = 1$. The specificity of $M_2$ relative to $M_1$ is defined as the value of the minimum of $\Psi$:

$$\Phi_T(\Delta_0^*; s_T) = \mathbb{E}_P[\Psi(\Delta_0^*; s_T)]$$

A number of "distance" criteria exist, among which we can usefully consider the following:

(i) The Kullback-Leibler information criterion:

$$\Psi_K(\Delta; s_T) = \Sigma_i \log \left[ \frac{\nu^*(j|s_T)}{\nu(j|s_T)} \right] \nu^*(j|s_T)$$

(ii) The Hellinger distance:

$$\Psi_H(\Delta; s_T) = \Sigma_i \left\{ [\nu^*(j|s_T)]^{\frac{1}{2}} - [\nu(j|s_T)]^{\frac{1}{2}} \right\}^2$$

(iii) The variation distance:

$$\Psi_1(\Delta; s_T) = \Sigma_i |\nu^*(j|s_T) - \nu(j|s_T)|$$

(iv) The $L^2$-norm:

$$\Psi_2(\Delta; s_T) = \Sigma_i [\nu^*(j|s_T) - \nu(j|s_T)]^2$$

The first three expressions are divergences - in the sense of Csiszár (1967) - between the posterior densities $\nu_\Delta^*$ and $\nu_T$ and, in particular, do not depend on the dominant measure on $B$. Furthermore (iii) is a genuine distance.

We have been unable to obtain analytical expressions for $\Delta_0^*$ under the first three criteria. Their numerical evaluation would be straightforward. Approximations are discussed below. Under the $L^2$-norm criterion we can obtain an analytical expression for $\Delta_0^*$, subject to the adding-up constraint ($\Sigma_i \delta_{ij} = 1$) at the cost of deleting the positivity constraints ($\delta_{ij} \geq 0$). As shown in theorem 2 below, the latter are satisfied at the limit ($T \rightarrow \infty$) and, in any event, they can easily be implemented numerically for application purposes.

**Theorem 2:** Under the assumptions of section 2.1, if $M_1$ is identified and if $E_P(\mu_T \mu_T')$ is non singular, then the pseudo-true value $\Delta$ which minimizes expression (3.5) under the $L^2$-norm (3.10) and subject to the adding up constraint $\Sigma_i \delta_{ij} = 1$ is given by

$$\Delta_0^* = \left[ E_P(\mu_T \mu_T') \right]^{-1} E_P(\mu_T \nu_T') (I - \frac{1}{n} \epsilon' \epsilon) + \frac{1}{n} \epsilon' \epsilon'$$

---

1 The arg. notation is convenient for expository purposes. In the case of multiple solutions, it implicitly requires that the choice of a specific solution is made unambigious. The notion of specificity in formula (3.8) is naturally invariant relative to that choice.
where $e' = (1, \ldots, 1)eR^n$.

If, in addition, assumption $H$ in (2.7) applies, then

$$
\lim_{T \to \infty} \delta^0_T(i,j) = d_{ij}
$$

(3.12)

where $d_{ij}$ is defined in formula (2.3).

**Proof:** See Appendix.

We conclude this section by briefly discussing (i) the extent to which our analysis can accommodate cases where $M_1$ is unidentified and, (ii) the relationship between the expression of $\Delta^0_T$ given in (3.11) and that of an approximation which has been proposed by Florens et. al. (1989) in a broader context.

There is a close relationship between the non-singularity of $E_P(\mu \mu_T^T)$ and the identification of $M_1$. If, in particular, $i$ and $i'$ are observationally equivalent, then the corresponding rows (columns) in $E_P(\mu \mu_T^T)$ are proportional to each other and $E_P(\mu \mu_T^T)$ is singular. It follows that the minimization problems (3.5) has multiple solutions though, obviously, the specificity (3.6) remains uniquely defined. As an addenda to the proof of theorem 2 in Appendix we also demonstrate that if $M_1$ is identified then $E_P(\mu \mu_T^T)$ is non-singular, at least for large enough $T$'s.

We chose to derive our main results under the (implicit) assumption that $M_1$ is identified largely for the ease of presentation. The obvious way to generalize our analysis to situations where $M_1$ is not identified consists in replacing $M_1$ by an identified "canonical" $\tilde{M}_1$ defined as follows. Let $\tilde{A}$ denote the quotient space of $A$ by the observational equivalence relationship:

$$
i \sim i' \iff f(x_i|i) = f(x_i'|i') \quad a.s.
$$

(3.13)

The model $\tilde{M}_1$, with parameter space $\tilde{A}$, is characterized by the following sampling probabilities:

$$
\tilde{p}(s_T|i) = p(s_T|i) \quad \forall i \tilde{i}
$$

(3.14)

where $\tilde{i}$ is a short hand notation for $A_i$, the equivalence class of $i$. The prior probability $\tilde{\mu}$ on $\tilde{A}$ is defined by the trace of $\mu$ on $\tilde{A}$:

$$
\tilde{\mu}(\tilde{i}) = \sum_{i \in \tilde{i}} \mu(i)
$$

(3.15)

The replacement of $M_1$ by $\tilde{M}_1$, finds its justification in the following lemma:

**Lemma 3:** The specificity of $M_2$ relative to $M_1$ is the same as that of $M_2$ relative to $\tilde{M}_1$.

**Proof:** See Appendix.

In their discussion of encompassing Florens et. al. (1989) note that analytical solutions to the minimization problem (3.5) may not be available and, therefore, propose alternative choices to $\Delta^0_T$ which, though they are not optimal, have reasonable heuristic interpretations and appropriate asymptotic behavior. One such transition $\tilde{\Delta}_T = (\tilde{\delta}_T(i,j))$ is the sampling expectation of the posterior probability $\nu_T$ associated with $M_2$:

$$
\tilde{\delta}_T(i,j) = \int_{S_T} \nu(j|s_T)p(s_T|i)ds_T
$$

(3.16)

In the case under consideration $\tilde{\Delta}_T$ takes an especially simple form and is given by

$$
\tilde{\Delta}_T = D^{-1}_{\mu} E_P(\mu \nu_T^T)
$$

(3.17)

where $D_{\mu}$ is diagonal matrix whose $i$-th diagonal element is $\mu(i)$. By comparison with formula (3.11), we see that $\tilde{\Delta}_T$ differs from $\Delta^0_T$ in that $E_P(\mu \mu_T^T)$ is replaced by $D_{\mu}$ which, as shown in section 4 below, happens to be its large sample limit (Also the adding-up factors in (3.11) are deleted but they are anyway negligible for large $T$'s). Hence, in the present context, $\Delta_T$ is a sensible approximation to $\Delta^0_T$ and, furthermore, has the same limit.

**Lemma 4:** Under the conditions of theorem 2, including assumption $H$, we have

$$
\lim_{T \to \infty} \tilde{\delta}_T(i,j) = d_{ij}
$$

(3.18)
4. Asymptotic Encompassing

In section 3, we have obtained an analytical expression for the optimal transition $\Delta_0^T$ associated with the $L^2$-norm criterion and have used that expression to demonstrate that $\Delta_0^T$ lends to $D$, as defined in (2.3).

Let us first extend this result by examining the asymptotic behavior of the criterion function $\Psi_2(\Delta_0^T; s_T)$ and of its expression $\Phi_T(\Delta_0^T)$, as defined in formulae (3.10) and (3.6) respectively.

**Theorem 5:** Under the assumptions of theorem 2, including $H$, we have:

(i) $\Psi_2(\Delta_0^T, s_T) \to 0$, $P^i$ a.s. \hspace{1cm} (4.1)

(ii) $\Phi_T(\Delta_0^T) \to 0$ \hspace{1cm} (4.2)

**Proof:** See Appendix.

In words, as $T$ gets arbitrarily large, a "true" model encompasses all its rivals, in the sense that the specificities of the latter relative to the former tend to zero. Formula (4.1) indicates that this limit result also applies to the "conditional specificity" $\Psi_2(\Delta_0^T; s_T)$ which therefore could be used as a statistic for testing the validity of $M_1$ (relative to $M_2$). Bayesian tests are extensively discussed by Florens and Mouchart (1988). In order to apply their line of reasoning to our problem, we would have to derive the distribution of that statistic - or of suitable approximations thereof - under $M_1$ and also under the alternative $M_2$ (i.e. either under the sampling distribution $Q$ or under the joint distribution $\chi$). This analysis goes beyond the objective of our paper.

In summary, $\Delta_0^T$ exhibits two important asymptotic properties:

(i) It converges towards the (classical) pseudo-true value $D$, as defined in formula (2.3);

(ii) Both the conditional and unconditional specifications of $M_1$ relative to $M_2$ tend to zero when $M_1$ is a "valid" model.

Our proof of theorem 5 makes use of the analytical expression of $\Delta_0^T$, as given in formula (3.11). In fact, theorem 5 can be generalized in order to cover the use of other criterion functions, such as those we introduced in formula (3.7) - (3.9), when no analytical expressions are available for the optimal transition. Such extensions are based upon a technique that is commonly used for the derivation of asymptotic properties of estimators and which essentially consists in permuting the optimization and limit operators. See e.g. Amemiya (1985). Adapted to the context of the problem under discussion, the corresponding theorem takes the following form:

**Theorem 6:** Let $\Phi_T(\Delta)$ denote a criterion function defined on $\mathcal{C}$, the compact set of $m \times n$ transition matrices. If:

(i) $\Phi_T(\Delta)$ is continuous in $\Delta$ and has a unique minimizer $\Delta_0^T$;

(ii) $\Phi_T(\Delta)$ uniformly converges on $\mathcal{C}$ towards a function $\Phi(\Delta)$ which has a unique minimizer $\Delta_0$;

Then

(i) $\lim_{T \to \infty} \Delta_0^T = \Delta_0$ \hspace{1cm} (4.3)

(ii) $\lim_{T \to \infty} \Phi_T(\Delta_0^T) = \Phi(\Delta_0)$ \hspace{1cm} (4.4)

If, furthermore, $\Phi_T(\Delta)$ is itself the expectation of a function $\Phi(\Delta; s_T)$, which converges $P^i$ almost surely and uniformly in $\Delta$ towards a function $F(\Delta; i)$, then

(iii) $\lim_{T \to \infty} \Psi(\Delta_0^T, s_T) = F(\Delta_0; i)$ \hspace{1cm} $P^i$-almost surely \hspace{1cm} (4.5)

**Proof:** See Appendix.

The application of theorem 6 to the criterion function (3.7) - (3.9) and to their expectation requires the following steps:
(1.) Using the asymptotic results in theorem 1, we first have to establish that the \( \mathcal{F}_i \)-almost sure limit of \( \Psi(\Delta; s_T) \) is a function \( F(\Delta; i) \) which does not depend on the sample sequence. Hence \( \Psi(\Delta; s_T) \) converges \( \Pi \)-almost surely towards \( F(\Delta; i) \). After verification that the latter also implies convergence of the expectation, we find that:

\[
\Phi_T(\Delta) = E_F[\Psi(\Delta; s_T)] = E_{\Pi}[\Psi(\Delta; s_T)] \\
= E_{\Pi}[F(\Delta; i)] = E_F[F(\Delta; i)] = \Sigma_i \mu(i) F(\Delta; i) = d_{ij} \Psi(\Delta) \quad (4.6)
\]

(2.) We next establish that \( \Phi(D) = 0 \), where \( D \) is the transition defined in (2.3). Hence, \( \Delta_0 = D \) since \( \Phi(\Delta) \geq 0 \).

(3.) Finally we can prove that \( F(D; i) = 0 \).

Formula (4.3) - (4.5) then follow by application of Theorem 6.

For the purpose of illustration we conclude this section with an application of this general technique to the Hellinger distance criterion function (3.7). Following formula (3.3) and condition (i) in theorem 1, we have

\[
F(\Delta; i) = \Sigma_j \left\{ \Sigma_{i'j} \mu(i'|Ai) \delta_{ij} \right\}^2
\]

This expression is bounded above by 2. Hence, the dominated convergence theorem applies and \( \Phi(\Delta) \) is given by \( \Sigma_i \mu(i) F(\Delta; i) \). If, furthermore, \( M_1 \) is identified, then \( A_i = \{ i \} \) and \( \mu(i'|i) = 1 \) for \( i = i' \) and is zero otherwise. Hence

\[
\Phi(\Delta) = \Sigma_{ij} \mu(i) (\delta_{ij}^\frac{1}{2} - \delta_{ij})^2
\]

By assumption \( \mu(i) > 0 \), \( \forall i \). Hence \( D \) is the unique minimizer of \( \Phi(\Delta) \) and \( \Phi(D) = 0 \). A similar argument leads to the conclusion that \( F(D; i) = 0 \), \( \forall i \). The uniform convergence of \( \Phi_T(\Delta) \) follows from the fact that

\[
| \Sigma_{i'j} \mu(i'|s_T) \delta_{ij} - \delta_{ij} | = | \Sigma_{i'j} \mu(i'|s_T) (\delta_{ij} - \delta_{ij}) | \\
\leq \Sigma_{i'j} \mu(i'|s_T) \rightarrow 0 \quad P^i \text{a.s.}
\]

Hence \( \Sigma_i \mu(i'|s_T) \delta_{ij} \) converges uniformly towards \( \delta_{ij} \), \( P^i \) almost surely in \( \delta_{ij} \) and the uniform convergence of \( \Phi_T(\Delta) \) follows from the continuity of \( \Psi(\Delta^2; s_T) \).
Appendix: Proofs:

A.1 Proof of Theorem 1

(i) The proof of part (i) is standard and is often attributed to Doob. It generally follows from martingale convergence theorems. In the case of finite parametric spaces, direct proofs based on the strong law of large numbers are available. See e.g. DeGroot (1970), Freidman (1963,1965), or, for a more general analysis, Doob (1953), Hartigan (1983), Locarni (1980) and Florens et. al. (1988). There is no need to repeat these proofs here since, furthermore, (i) follows from (ii) if we consider identical models.

(ii) We proceed by generalizing the proof in DeGroot (1970) to “misspecified” models. The posterior density \( \nu(j|s_T) \) is obtained by application of Bayes theorem and may be rewritten as:

\[
\nu(j|s_T) = \left[ \sum_{\ell B_i} \frac{\nu(\ell)}{\nu(j)} \frac{q(s_T|\ell)}{q(s_T|j)} \right]^{-1} \tag{A.1}
\]

Consider first \( j \notin B_i \) and select an arbitrary \( \ell \) in \( B_i \), which is not empty by definition. For such a pair \((\ell, j)\) we find that:

\[
\frac{1}{T} \sum_{t=1}^{T} \log \left[ \frac{g(x_t|\ell)}{g(x_t|j)} \right] \Rightarrow \int \log \left[ \frac{f(x_t|i)}{g(x_t|j)} \right] f(x_t|i) \, dx_t - \int \log \left[ \frac{f(x_t|i)}{g(x_t|\ell)} \right] f(x_t|i) \, dx_t \tag{A.2}
\]

\( P^i \) almost surely by application of the strong law of large numbers. This limit is strictly positive by definition of \( B_i \). Hence

\[
\frac{q(s_T|\ell)}{q(s_T|j)} \to \infty \quad \text{and} \quad \nu(j|s_T) \to 0, \quad P^i \text{ almost surely}
\]

and (ii.b) follows from the fact that the posterior probabilities add-up to one.

(iii) Consider next \( j \in B_i \) and \( \ell \notin B_i \). For such a pair, the limit in (A.2) is strictly negative and \([q(s_T|\ell)/q(s_T|j)] \to 0, \ P^i\text{-almost surely. If instead } \ell \in B_i, \text{ then assumption } H \text{ in (2.7) implies that } [q(s_T|\ell)/q(s_T|j)] = 1 \) It follows that

\[
\nu(j|s_T) = \left[ \sum_{\ell \in B_i} \frac{\nu(\ell)}{\nu(j)} \right]^{-1} = d_{ij}, \quad P^i \text{ almost surely}
\]

and part (iii) follows from formula (2.3). \( \Box \)

A.2 Proof of Theorem 2

Formula (3.11) follows from the application of Lagrange multipliers techniques. Let us next examin the asymptotic behavior of \([E_{\Pi}(\mu_T \mu_T')^{-1} E_{\Pi}(\mu_T' \mu_T)]\), considering separately the limits of its two components.

By application of theorem 1 and since \( M_1 \) is identified, we find that \( \mu(i'|s_T) \to \mu(i'|i) \), where \( \mu(i'|i) = 1 \) if \( i' = i \) and equals zero otherwise. This convergence is \( P^i \)-almost sure since \( \mu(i) > 0, \forall i \). Hence

\[
\mu(i'|s_T) \mu(i''|s_T) \to \mu(i'|i) \mu(i''|i) \quad \Pi - a.s., \forall i', i''
\]

Since these probabilities are bounded, the dominated convergence theorem applies and

\[
E_\Pi[\mu(i'|s_T) \mu(i''|s_T)] \to E_\Pi[\mu(i'|i) \mu(i''|i)] \quad \forall i', i''
\]

or, equivalently,

\[
E_P[\mu(i'|s_T) \mu(i''|s_T)] \to E_P[\mu(i'|i) \mu(i''|i)] \quad \forall i', i''
\]
since $\mu(i'|s_T)$ and $\mu(i'|i)$ depend solely on $s_T$ and $i$, respectively. The latter expectation is given by

$$E_\mu(i'|i)\mu(i''|i) = \delta_{ij}$$

Hence

$$E_P(\mu_T\mu_T^T) \rightarrow D_\mu$$

(A.3)

where $D_\mu$ is a diagonal matrix whose $i$-th diagonal element is $\mu(i)$. The product $\mu(i'|s_T)\nu(j|s_T)$ converges towards $\mu(i'|i)d_{ij}$ by application of theorem 1. Hence, by application of the dominated convergence theorem, we have

$$E_P(\mu_T\nu_T^T) \rightarrow D_\mu D$$

(A.4)

Formula (3.12) follows from (A.3) and (A.4) together with the fact that $D\epsilon = \epsilon 1$.

Remark: When $M_1$ is not identified $\Delta^0_T$ generally does not converge towards $D$. In fact $E_P(\mu_T\mu_T^T)$ then converges towards a block diagonal matrix whose $(i', i')$-th element is $\mu(i')\mu(i')/\mu(A_i)$ if $i$ and $i'$ are equivalent ($i' \epsilon A_i$) and zero otherwise. Hence $E_P(\mu_T\mu_T^T)$ is singular in the limit. $E_P(\mu_T\nu_T^T)$ converges towards $D_\mu D$ as before. Hence, even if we replace the inverse in formula (3.11) by a generalized inverse, the limit of $\Delta^0_T$ need not be $D$. In any event, the optimal transition $\Delta^0_T$ associated with the identified canonical model $M_1$ defined in (3.14) does converge towards a matrix $\tilde{D}$, which is derived from $D$ by deletion of the rows corresponding to all but one of the equivalent elements.

A.3 Proof of Lemma 3

It is sufficient to demonstrate that the set of "derived" posterior densities for $j$, as defined in (3.3), is unaffected by the replacement of $M_1$ by $\tilde{M}_1$. Let

$$E = \{\nu^*_T|\nu^*(j|s_T) = \delta_{ij}, \delta_{ij} \geq 0, \sum_i \delta_{ij} = 1\}$$

$$\tilde{E} = \{\tilde{\nu}_T^*|\tilde{\nu}^*(j|s_T) = \delta_{ij}, \delta_{ij} \geq 0, \sum_i \delta_{ij} = 1\}$$

(i) Let $\tilde{\nu}_T^* \epsilon \tilde{E}$ and let $\delta_{ij} = \delta_{ij}, \forall i \tilde{i}$. Then

$$\tilde{\nu}^*(j|s_T) = \delta_{ij} = \sum_i \mu(i|s_T)\delta_{ij} = \sum_i \mu(i|s_T)\delta_{ij}$$

and $\tilde{E} \epsilon E$.

(ii) Let $\nu_T^* \epsilon E$. By definition of $\tilde{i}$, we have that $\mu(i|\tilde{i}, s_T) = \mu(i|\tilde{i})$. Hence

$$\nu^*(j|s_T) = \sum_i \mu(i|s_T)(\sum_i \mu(i|\tilde{i})\delta_{ij}) = \sum_i \mu(i|s_T)\delta_{ij}$$

where $\delta_{ij} = \sum_i \mu(i|\tilde{i})d_{ij}$ and $E \epsilon \tilde{E}$.

A. 4 Proof of Theorem 5

(i) $\Psi_T(\Delta^0_T; s_T) = (\Delta^0_T\mu_T - \nu_T)^T(\Delta^0_T\mu_T - \nu_T)$. In the course of our proof of theorem 2 we have established that $\Delta^0_T \rightarrow D$, $\mu_T(i') \rightarrow \mu(i'|i)$ and $\nu_T(j) \rightarrow d_{ij}$, $P^i$ almost surely. Hence $\Delta^0_T\mu_T \rightarrow d_{ij}$, $P^i$ almost surely and formula (4.1) follows.

(ii) $\Phi_T(\Delta^0_T) = tr[\Delta^0_T E_P(\mu_T\mu_T^T)\Delta^0_T - 2\Delta^0_T E_P(\mu_T\nu_T^T) + E_P(\nu_T\nu_T^T)]$. An immediate extension of our proof of formula (A.4) leads to the conclusion that $E_P(\nu_T\nu_T^T) \rightarrow D^T D$ and formula (4.2) follows from (A.3) and (A.4).
A. 5 Proof of Theorem 6

\( C \), the set of \( m \times n \) transition matrices is endowed with the induced Euclidian topology. We note in passing that the compactness of \( C \) together with the continuity of \( \Phi_T(\Delta) \) and \( \Phi(\Delta) \) guarantees the existence of \( \Delta^0_\frac{1}{2} \) and \( \Delta^0 \). Let \( B \) denote an arbitrary open neighborhood of \( \Delta_0 \). In order to establish formula (4.3), we verify that

\[
\exists T^*, \quad T \geq T^* \to \Phi_T(\Delta) > \Phi_T(\Delta_0), \quad \forall \Delta \epsilon C \setminus B,
\]

where \( a = \Phi(\Delta) - \Phi(\Delta^0) \), \( b = \Phi(\Delta^0) - \Phi_T(\Delta^0) \) and \( c = \Phi_T(\Delta) - \Phi(\Delta) \). For any \( \Delta \epsilon C \setminus B \) we have \( a \geq \alpha \), where \( \alpha = \inf_{C \setminus B} [\Phi(\Delta) - \Phi(\Delta^0)] > 0 \). By the uniform convergence of \( \Phi_T \), we can select \( T^* \) in such a way that

\[
|\Phi_T(\Delta) - \Phi(\Delta)| < \frac{1}{2}\alpha \quad \text{for any} \quad \Delta \epsilon B \quad \text{and} \quad \text{formula (4.5) follows.}
\]

Formula (4.4) follows from the triangular inequality

\[
|\Phi_T(\Delta^0_\frac{1}{2}) - \Phi(\Delta^0)| \leq |\Phi_T(\Delta^0_\frac{1}{2}) - \Phi(\Delta^0_\frac{1}{2})| + |\Phi(\Delta^0_\frac{1}{2}) - \Phi(\Delta^0)|
\]

where the first factor in the r.h.s of the inequality tends to zero by uniform convergence and so does the second one by the continuity of \( \Phi \), which follows from the continuity of \( \Phi_T \) and from the uniform convergence assumption. \( \square \)
List of References


