Bayesian Analysis and Model Revision for k'th Order Markov Chains with Unknown k. [Preliminary Draft, October 2, 2002]

by

Roy Radner

Stern School of Business

New York University

44 West Fourth Street

New York, NY 10012

Abstract

Consider the problem of consistent Bayesian estimation of a stationary "k'th-order Markov process" on a finite state space, when the parameter k is itself unknown, as well as the transition probabilities for each value of k. First, I show that if k has a known upper bound, then on a single realization of the process the posterior probability measures on the parameter space converge weakly to a probability measure with

Key Words and Phrases: Bayesian analysis, model revision, bounded rationality

mass 1 concentrated on the true process, provided that the prior probability measure has full support and the true process is irreducible. Second, I extend this result to the case in which k is unbounded (but finite), which requires that the Bayesian decision-maker (DM) construct a prior on an infinite-dimensional parameter space. Finally, in an alternative approach to this case, I suppose that the DM considers a succession of models corresponding to larger and larger values of k. Each time the DM revises his model he extends his prior probability measure to the new - and larger - parameter space in a way that is "consistent" with the previous prior, and recomputes his posterior probability measures. I show that, roughly speaking, if the DM does not revise his model “too frequently,” then he will be increasingly confident that the current posterior is increasingly concentrated on the true process. I motivate the procedure of model revision by considerations of *bounded rationality.*

1 Introduction

In these notes I consider the problem of consistent Bayesian estimation of a stationary “k'th-order Markov process” on a finite state space, from a single realization of the process, when the parameter k is itself unknown, as well as the transition probabilities for each value of k. (Recall that, roughly speaking, a k'th-order Markov process is a stochastic process such that, for each date, the conditional probability distribution of the future given the past depends at most on the last k observations. A more precise definition is given below.) Since the parameter space for this problem is infinite-dimensional, consistent Bayesian estimation may be problematic. At least, such a concern is suggested by the analogous case of estimating an infinite-dimensional parameter vector from a sequence of independent and identically distributed random variables (see Freedman, 1963, and Diaconis and Freedman, 1986). The case considered here differs from the latter in two respects. On the one hand, the successive observations are assumed to take values in a finite set. On the other hand, the successive observations are not assumed to be mutually independent.

Let $M(k)$ denote the space of stationary k'th-order Markov processes on a given finite state space. Note that if $k' \leq k$, then $M(k')$ can be identified with a subset of $M(k)$ in a natural way. I first consider the case in which k has a known upper bound. In this case the parameter space of the process is finite-dimensional, and one can use Freedman's method to show that the posterior probability measure on the parameter space converges weakly to a probability measure with mass 1 concentrated on the true process, provided that the prior probability measure has full support and the true process is irreducible (Theorem 6).

I next consider the case in which k is unbounded (but finite). Let M denote the union of the sets $M(k)$. I show that the conclusion of Theorem 6 remains true if the prior on M is a convex combination of priors on $M(k)$ for which the assumptions of Theorem 6 are satisfied for all sufficiently large k (Theorem 9).

In the case covered by Theorem 9, in which an upper bound on the parameter k is not known, the Bayesian must construct a prior on the entire space M. Since M is infinite dimensional, it might be difficult for an individual to formulate beliefs about such a large space that are sufficiently precise to yield a meaningful prior. In the last section I outline an alternative approach, which is a process of *model revision.* Suppose that the Bayesian decision-maker (DM) considers a succession of models corresponding to larger and larger values of k. Each time the DM revises his model he extends his prior probability measure to the new - and larger - parameter space in a way that is "consistent" (in a way to be specified) with the previous prior, and recomputes his posterior probability measures. I show that, roughly speaking, *if the DM does not revise his model "too frequently," then he will be increasingly confident that the current posterior is increasingly concentrated on the true process.* Indeed, since the dimension of the parameter space increases geometrically with k, the time intervals between successive model revisions must be increasingly long.

The procedure of model revision studied here can be motivated by considerations of *bounded rationality.* At the beginning of an empirical investigation, a DM will typically find it difficult to formulate beliefs about a very large parameter space that are sufficiently precise to yield a meaningful prior probability measure on the space. Savage (1954, pp. 59, 168, 169) refers to this as the "problem of vagueness." (See also Radner, 1996, 1999). However, as he gains experience with the investigation the DM may explore increasingly complex models with increasingly large parameter spaces. In this context, a natural question is whether there is a "rational" way to do this. The present analysis can be seen as a step in trying to answer this question.

Acknowledgements

I am grateful to David Freedman and Hsueh-Ling Huang for helpful discussions of this topic, and to Shmuel Zamir for comments on an earlier draft.

2 Preliminaries

In this section I formulate the model and present some preliminalry results that will be useful tools for the analysis of the following sections. These results are standard and/or can be derived using standard methods, but are gathered here for the convenience of the reader. With an eye to possible generalizations and extensions of the main results, I have presented a formulation that is somewhat more general than is strictly required for what follows.

Let X denote the set of all infinite sequences, $\{x(t)\}$, where t ranges over $\mathbb{N}$, the nonnegative integers, and $x(t)$ takes values in some finite set, I, the same for all t. Call I the *state space*. For concreteness, think of t as time (discrete), and I as a finite set of integers. Let every subset of I be measurable, and let H denote the product sigma-field of measurable subsets of X. For each integer t, let $H(t)$ denote the subsigma-field generated by "histories" up through date t. For short, the measurable space (X, H) will be denoted by X. The space X can also be metrized in a compatible way. Take the distance between two integers in I to be the absolute value of their difference, and endow X with the corresponding product topology. This is also the metric topology with the metric ,

$$\sigma(x, y) = \sum_{t=0}^{\infty} \left(\frac{1}{2^t}\right) |x(t) - y(t)|, \text{ for } x, y \text{ in } X. \tag{1}$$

For this topology, X is a compact metric space, and the Borel sets of X are the sets in H. Note that, since I is finite,

$$\lim_{n\to\infty} x_n = x \iff \text{for every } t, \text{ eventually } x_n(t) = x(t). \tag{2}$$

Let $P(X)$ denote the set of (countably additive) probability measures on X, or, in a different language, the set of stochastic processes with state space I. [More generally, if $(Y, \mathcal{Y})$ is a measurable space, let $P(Y)$ denote the set of probability measures on $(Y, \mathcal{Y})$.] Let $C(X)$ denote the set of bounded continuous real-valued functions on X. Endow $P(X)$ with the topology of *weak* convergence*, namely, for $\{m_n\}$ and m in $P(X)$:

$$m_n \to m \equiv \int c(x)dm_n(x) \to \int c(x)dm(x), \text{ for every } c \in C(X).$$

Since X is a compact metric space, so is $P(X)$ (see Parthasarathy, 1967, Ch. II, Theorem 6.4, p. 45).

Proposition 2 below gives another characterization of weak* convergence in $P(X)$. Proposition 3 describes a base for the open sets in $P(X)$. Endow $C(X)$ with the sup norm topology. The first step in this analysis is to construct a countable dense subset of $C(X)$. For each $n \geq 0$ let F_n denote the set of all real-valued functions on X that depend only on the first $n+1$ coordinates, $x(0), ..., x(n)$. Note that, by (2), every element of F_n is in $C(X)$. Also, F_n is essentially $R^{(n+1)I}$. Let Γ_n be a countable dense subset of F_n, and define the countable set Γ by:

$$\Gamma = \cup_{n \geq 0} \Gamma_n. \tag{3}$$

Proposition 1 Γ *is dense in* $C(X)$.

Proof. Fix an element of I; call it 0. For each nonnegative integer n define the mapping ϕ_n from X into itself by:

$$\begin{aligned} \phi_n(x)(t) &= x(t), \text{ for } 0 \leq t \leq n, \\ &= 0, \text{ for } t > n. \end{aligned} \tag{4}$$

By (2),

$$\lim_{n \to \infty} \phi_n(x) = x, \text{ uniformly in } x. \tag{5}$$

Consider an $f \in C(X)$, and define $f_n \in F_n$ by

$$f_n(x) = f[\phi_n(x)].$$

Since f is uniformly continuous, for every $\epsilon > 0$ there is an n such that

$$|f_n(x) - f(x)| \leq \epsilon/2, \text{ for all } x \in X.$$

Also, there exists $g \in \Gamma_n$ such that

$$|g(x) - f_n(x)| \leq \epsilon/2, \text{ for all } x \in X.$$

Putting the last two statements together, we have: for every $\epsilon > 0$ there exists an n and a $g \in \Gamma_n$ such that

$$|g(x) - f(x)| \leq \epsilon, \text{ for all } x \in X.$$

Hence Γ is dense in $C(X)$, which completes the proof.

The next proposition provides an alternative characterization of weak* convergence in $P(X)$. For any nonnegative integer h and any subset of A' of I^h, the set

$$A = A' \times I^{\infty} \tag{6}$$

is called a *finite-dimensional cylinder set (of dimension* $h+1$*)*. Thus whether $x \in A$ depends only on the first h coordinates, $x(0), ..., x(h-1)$.

Proposition 2 *$m_n \to m$ in $P(X)$ if and only if $m_n(A)$ converges to $m(A)$ for every finite-dimensional cylinder set A.*

Proof. First suppose that $m_n \to m$. Let A be a finite-dimensional cylinder set, and let f be its indicator function, i.e., $f(x) = 1$ if $x \in A$ and 0 otherwise. Then $f \in F_h$ for some h (see the paragraph preceding Proposition 1), and hence is also in $C(X)$, so that

$$m_n(A) = \int_X f dm_n \to \int_X f dm = m(A).$$

Conversely, suppose that $m_n(A) \to m(A)$ for every finite-dimensional cylinder set A. By Proposition 1, for any $f \in C(X)$ there exists a sequence $\{f_j\}$ of functions in Γ such that

$$\|f_j - f)\| \to 0 \text{ as } j \to \infty.$$

Fix j, and observe that $j \in \Gamma_h$ for some h. For the purpose of this proof, let $\mathcal{I} = I^{h+1}$, and for any $a \in \mathcal{I}$ define

$$\begin{aligned} A(a) &= \{x \in X : x(t) = a(t)\} \text{ for } 0 \le t \le h, \\ \psi(a) &= f_j(x) \text{ for } x \in A(a); \end{aligned}$$

then

$$\int_X f_j dm_n = \sum_{a \in \mathcal{I}} \psi(a) m_n[A(a)] \to \sum_{a \in \mathcal{I}} \psi(a) m[A(a)] = \int_X f_j dm. \tag{7}$$

For any $\mu \in P(X)$,

$$\left| \int_X f d\mu - \int_X f_j d\mu \right| \le \int_X |f - f_j| \, d\mu \le \|f - f_j\|, \tag{8}$$

where $|f - f_j|$ denotes the function with values $|f(x) - f_j(x)|$. For every $\epsilon > 0$ there exists a $j(\epsilon)$such that

$$\begin{aligned} j &\ge j(\epsilon) \Rightarrow \|f - f_j\| \le \epsilon \Rightarrow \\ \left| \int_X f dm_n - \int_X f dm \right| &\le \left| \int_X f dm_n - \int_X f_j dm_n \right| + \left| \int_X f_j dm_n - \int_X f_j dm \right| \\ &\quad + \left| \int_X f_j dm - \int_X f dm \right| \\ &\le \left| \int_X f_j dm_n - \int_X f_j dm \right| + 2\epsilon. \end{aligned}$$

Hence, by (8), for every $\epsilon > 0$,

$$\limsup_{m\to\infty} \left| \int_X f dm_n - \int_X f dm \right| \leq 2\epsilon,$$

and hence

$$\lim_{m\to\infty} \left| \int_X f dm_n - \int_X f dm \right| = 0,$$

which completes the proof.

I next describe a base for the open sets of $P(X)$. Following Parthasarathy, 1967, Ch. 6, Theorem 6.2, p. 43, I shall construct a metric on $P(X)$ that generates the topology of weak* convergence. Let $\{g^j\}$ be a (countably infinite) listing of the elements of Γ. (Every g^j is a real-valued function on X that depends only on a finite number of coordinates; see Proposition 1.). Define the map J from $P(X) \times \mathbb{N}$ into R^∞ by

$$J(m, j) = \int g^j dm, \qquad j = 0, 1, 2, ...;$$

then J is a homeomorphism. Recall that convergence in R^∞ is pointwise convergence. Hence:

Proposition 3 *A base for the open sets of $P(X)$ is the family of all sets of the form*

$$B(E, \mathcal{A}) = \{m \in P(X) : J(m, j) \in A_j, \quad j \in E\}, \tag{9}$$

where E is a finite set of integers, for each j, A_j is an open subset of reals, and $\mathcal{A}$ is the family of sets $\{A_j : j \in E\}$.

I now consider $\mathcal{P} = P[P(X)]$, again with the topology of weak* convergence. In particular, I shall be interested in when a sequence $\{p_n\}$ in $\mathcal{P}$ converges to a probability measure concentrated on a single point, say m, in $P(X)$. The latter is sometimes called *the Dirac measure at* m, and will be denoted by $D(m)$. In the application, $\{p_n\}$ will be a sequence of "posterior probability measures" on $P(X)$, and m will be the "true" probability measure on X.

Proposition 4 *For* $m \in P(X)$,

$$\lim_{n \to \infty} p_n(A) = 1 \text{ for every open set } A \text{ containing } m$$

implies that

$$p_n \to D(m) \text{ in } \mathcal{P}.$$

Proof. Fix $m \in P(X)$, and for the purpose of the proof let $D = D(m)$. Let f be a real-valued, bounded, continous function on $P(X)$ (i.e., in $C(P(X))$), with bound b. First note that

$$\int_{P(X)} f dD = f(m).$$

Next, for any $\epsilon > 0$ there exists an open neighborhood A of m such that $q \in A \Rightarrow |f(q) - f(m)| \leq \epsilon$. Also, by the hypothesis, there exists an N such that

$$n \geq N \Rightarrow p_n(A) \geq 1 - \epsilon.$$

Hence, for $n \geq N$,

$$\begin{aligned}
\left|\int_{P(X)} f dp_n - \int_{P(X)} f dD\right| &= \left|\int_{P(X)} f dp_n - f(m)\right| \\
&= \left|\int_{P(X)} f dp_n - f(m)\int_{P(X)} dp_n\right| \\
&= \left|\int_{P(X)} [f(q) - f(m)] dp_n(q)\right| \\
&= \left|\int_{P(X)\setminus A} [f(q) - f(m)] dp_n(q)\right| + \left|\int_A [f(q) - f(m)] dp_n(q)\right| \\
&\leq 2b\epsilon + \epsilon.
\end{aligned}$$

Hence, for every $\epsilon > 0$,

$$\limsup_{n\to\infty} \left|\int_{P(X)} f dp_n - \int_{P(X)} f dD\right| \leq 2b\epsilon + \epsilon.$$

It follows that

$$\lim_{n\to\infty} \int_{P(X)} f dp_n = \int_{P(X)} f dD.$$

Since this is true for every such f, the conclusion of the proposition follows.

Now consider a particular $m' \in P(X)$, and let $D' = D(m')$ be the Dirac measure at m'. Let $\{p_n\}$ be a sequence of measures in $\mathcal{P}$. By Proposition 4, $p_n \to D'$ if

$$\lim_{n\to\infty} p_n[B(E, \mathcal{A})] = 1 \text{ for every set } B(E, \mathcal{A}) \text{ of the form (9) containing } m'. \qquad (10)$$

In these notes, I shall concentrate on a particular subset of $P(X)$. For a positive integer k, a *k'th-order Markov process* is a stochastic process on X such that, for each date $t \geq k-1$, the conditional probability distribution of the "future" after date t, conditioned on $H(t)$, depends only on the last k coordinates, $x(t), x(t-1), , x(t-k+1)$.

Recall that a stochastic process on X (i.e., a measure in $P(X)$) is *stationary* if it is invariant under translations in X (see, e.g., Doob, 1953, Ch. X). The corresponding measure in $P(X)$ is also called stationary. For each positive integer k, let $M(k)$ denote the set of *stationary* k'th-order Markov processes. Make the convention that $M(0)$ is the set of independent and identically distributed sequences taking values in I. Let M denote the union of all of the sets $M(k)$, for k nonnegative.

The following proposition is not used directly in the analysis of Bayesian inference in M, but it gives some perspective on how "large" M is. Let S denote the set of all stationary stochastic processes with state space I, i.e., the set of stationary measures in $P(X)$.

Proposition 5 *M is dense in S in the topology of weak* convergence in $P(X)$.*

(Proof omitted.)

It will be useful to parametrize M. For the moment, fix k. A k'th-order Markov process can be parametrized by a transition matrix, Q, and an initial probability vector, s, as follows. Any sequence $[x(t), ..., x(t+k-1)]$ takes values in the k-fold product of I with itself. A *transition matrix*, $Q = ((q_{hj}))$, describes the conditional probability that the next state is j, given that the immediately preceding history of

k observations is h. Thus,

$$\begin{aligned}
Q &= ((q_{hj})),\ h \in I^k, , j \in I; \\
\sum_{j \in I} q_{hj} &= 1,\ \text{for every } h \in I^k; \\
\Pr\{X(t+1) &= j \mid [X(t), ..., X(t+k-1)] = h\} = q_{hj},\ \text{for all } t \geq k-1.
\end{aligned} \tag{11}$$

An *initial probabiity vector,* (s_h), describes the joint probability distribution of $[X(0), ..., X(k-1)]$. Thus,

$$\begin{aligned}
s &= (s_h), h \in I^k; \\
\sum_h s_h &= 1; \\
\Pr\{[X(0), ..., X(k-1)] &= h\} = s_h.
\end{aligned} \tag{12}$$

A generic pair (Q, s) will be denoted by θ. One can identify $M(k)$ with the space $\Theta(k)$ of all possible such pairs, which is a subset of a closed, compact subset of a Euclidean space of dimension $|I|^{k+1} + |I|^k$, where $|I|$ denotes the number of elements of I. When there is no risk of confusion, I shall use the same symbol, $M(k)$, for the space of measures and the corresponding parameter space. Note that if $k' < k$, then $M(k')$ can be identified in a natural way with a lower dimensional manifold of $M(k)$.

Finally, note that, since I is finite, weak* convergence in $M(k)$ is equivalent to Euclidean convergence in $\Theta(k)$.

3 Bayesian Analysis in M

3.1 Bayesian Analysis in $M(k)$

Recall that M is the union of all the $M(k), k \geq 0$. We are interested in "prior" and "posterior" probability distributions on M. A *prior (probability distribution) on M* is a probability measure on M, i.e., an element of $P(M) \subset P[P(X)]$. A prior μ on M generates a probability measure on $M \times X$ as follows. A "parameter" θ is chosen according to μ. Conditional on θ, the stochastic process $X_\infty = \{X(t)\}$ is generated according to the k'th-order Markov process corresponding to θ. Let x be a point in X, let $X_T =\ [X(0), ..., X(T)]$ be the first $(T+1)$ coordinates of a particuar realization of the stochastic process X_∞, and let $L_T(x, \theta)$ be the conditional probability that $X_T = [x(0), ..., x(T)]$, given the parameter θ, i.e.,

$$L_T(x, \theta) = \Pr\{X_T = [x(0), ..., x(T)] | \theta\}. \tag{13}$$

L_T is called the *likelihood function* corresponding to the first $(T+1)$ observations.

Given a prior μ on M, a (measurable) set A of parameters in M, a point x in X, and a date $T \geq 0$, define

$$\psi(x, T)(A) \equiv \frac{\int_A L_T(x, \theta) d\mu(\theta)}{\int_M L_T(x, \theta) d\mu(\theta}, \tag{14}$$

provided the denominator is not zero. The probability measure $\psi(x, T)$ on M is called the *posterior distribution* corresponding to the first $(T+1)$ observations. Thus $\psi(x, T)(A)$ is the conditional probability that $\theta \in A$, given that $X_T = [x(0), ..., x(T)]$. (This statement is called *Bayes's Theorem.*)

Suppose now that we are given a particular θ' in M, and that the probability law of the stochastic process X_∞ is determined by θ'. For fixed T and A, the posterior measure of A, namely $\psi(X_\infty, T)(A)$, is a random variable, whose probability distribution is governed by θ'. In other words, $\psi(X_\infty, T)$ is a "random probability measure" on the parameter space M. We are interested in conditions under which this sequence of random measures, $\{\psi(X_\infty, T)\}$ almost surely converges weak* to a probability measure that is entirely concentrated on the point θ'. As noted above, the latter measure is sometimes called the *Dirac measure at* θ', and will be denoted by $D(\theta')$. In this case we shall call the pair (μ, θ') *consistent*, or say that μ *is consistent for* θ'. By Proposition 4, a sequence $\{m_T\}$ converges weak* to the Dirac measure at θ if for every open neighborhood B of θ the sequence $\{m_T(B)\}$ converges to unity.

For $\theta \in M(k)$, I shall say that θ is *irreducible* if the k'th-order Markov chain corresponding to θ is irreducible. Recall that, for any k, the weak* topology on $M(k)$ is equivalent to the the Euclidean topology on $\Theta(k)$.

Theorem 6 *If (1) μ is a prior on M such that $\mu[M(k)] = 1$, (2) θ' is in the support of μ, and (3) θ' is irreducible, then μ is consistent for θ'.*

The proof is based on Theorem 1 of (Freedman, 1963), and relies on the fact that $M(k)$ is finite dimensional. Fix k, and for the purpose of this proof let K denote the k-fold product of I with itself. Fix $\theta' = (Q', s')$ in $M(k)$; recall that $Q' = ((q'_{hj}))$, $h \in K$, $j \in I$, is the matrix of transition probabities, and $s' = (s'_h)$ is the vector of probabilities of the initial strings of length k. Since θ' is irreducible, s' is uniquely determined by Q'. In the rest of the proof (unless otherwise noted), all random variables and probabilities refer to the stochastic process, X_∞, generated by θ'. For any point h in K and any date t, I shall say that h *occurs at* t *if* $[X(t-k+1), ..., X(t)] = h$. Let $\{T_n(h)\}$ be the sequence of (random) times at which h occurs. The irreducibility of θ' implies that this sequence is a.s. infinite. For a given h, the random variables $X[T_n(h)+1]$ are IID, with values in I, and probability distribution given by row h of the transition probability matrix Q'.

Fix T, and consider the $(k+T)$ observations, $[X(0), ..., X(k+T-1)]$.Define

$$\begin{aligned} N_h &= \text{no. of occurences of } h, \\ N_{hj} &= \text{no. of times an occurence of } h \text{ is followed by state } j. \end{aligned} \tag{15}$$

Note that

$$\begin{aligned} \sum_{j \in I} N_{hj} &= N_h, \\ \sum_{h \in K} N_h &= T. \end{aligned} \tag{16}$$

The value of the likelihood function, (13), is

$$L_{k+T-1}(x, \theta) = s[x(0), ..., x(k-1)] \prod_{h \in K} \prod_{j \in I} q_{hj}^{N_{hj}}. \tag{17}$$

Taking the logarithm of the likelihood, we have

$$\ell_{k+T-1}(x, \theta) \equiv \ln L_{k+T-1}(x, \theta) = \ln s[x(0), ..., x(k-1)] + \sum_{h \in K} \sum_{j \in I} N_{hj} \ln q_{hj}.$$

Dividing this last by T, we have

$$\left(\frac{1}{T}\right)\ell_{k+T-1}(x,\theta) \equiv \left(\frac{1}{T}\right)\ln s[x(0),...,x(k-1)]+\sum_{h\in K}\left(\frac{N_h}{T}\right)\sum_{j\in I}\left(\frac{N_{hj}}{N_h}\right)\ln q_{hj}. \quad (18)$$

As T increases without bound, the following hold a.s. with respect to θ':

$$\begin{aligned} \lim_{T\to\infty} N_{hj} &= \lim_{T\to\infty} N_h = \infty, \\ \lim_{T\to\infty}\left(\frac{1}{T}\right)\ln s[x(0),...,x(k-1)] &= 0, \\ \lim_{T\to\infty}\frac{N_h}{T} &= s'(h), \\ \lim_{T\to\infty}\frac{N_{hj}}{N_h} &= q'_{hj}. \end{aligned} \quad (19)$$

In particular, the sums,

$$\sum_{j\in I}\left(\frac{N_{hj}}{N_h}\right)\ln q'_{hj}, \quad (20)$$

behave asymptotically as in the case of IID random variables, and there are only finitely many such sums. This suggests that the argument of Theorem 1 of (Freedman, 1963) can be modified to yield the desired conclusion. To carry this out, it will be useful to have at hand an explicit proof for the IID case.

Lemma 7 *(Freedman, 1963) Let $\{X(t))$ be independently and identically distributed (IID) with values in the finite set I. Let U denote the unit simplex in R^I. A probability measure on I is a point in U. Let μ be a probability measure on the Borel sets of S (the "prior") with full support, and let q' be a particular point in U such that $q' >> 0$ (strictly positive coordinates). Then μ is consistent for q'.*

Proof. For the reader's convenience, I specialize the notation of Section 3 to this case. Let $N_i(T)$ denote the number of dates t in the sample $[X(0), ..., X(T-1)]$ at which $X(t) = i$. Of course,

$$\sum_{i \in I} N_i(T) = T;$$

the *relative frequencies* in the sample are given by

$$F_i(T) = (1/T)N_i(T).$$

The *likelihood* of the sample, given the probabilty vector $q \in U$, is

$$L(T, q) = \prod_{i \in I} q_i^{N_i(T)}, \tag{21}$$

and the posterior probability of a (measurable) subset B of U is

$$\psi(\cdot, T)(B) = \frac{\int_B L(T, q) d\mu(q)}{\int_U L(T, q) d\mu(q)}. \tag{22}$$

The *loglikelihood* of the sample, $\ell(T, q)$, is the natural logarithm of the likelihood; thus

$$\begin{aligned} \ell(T, q) &= \ln L(T, q) = \sum_i N_i(T) \ln q_i, \\ (1/T)\ell(T, q) &= \sum_i F_i(T) \ln q_i. \end{aligned} \tag{23}$$

By the Strong Law of Large Numbers, if the $X(t)$ have the probability distribution q', then

$$\lim_{T \to \infty} (1/T)\ell(T, q) = \sum_i q_i' \ln q_i \text{ a.s.} \tag{24}$$

For y and z in U, define

$$G(y,z) = \sum_i y_i \ln z_i, \tag{25}$$

with the understanding that $0 \ln 0 = 0$. Note that G is continuous on U^2, and hence uniformly continuous. Also, for each $y, G(y,z)$ is concave in z, and strictly concave if $y >> 0$. Furthermore,

$$\max_{z \in U} G(y,z) = G(y,y), \tag{26}$$

and the maximum is attained uniquely at $z = y$ if $y >> 0$. Finally, note that

$$\begin{aligned} (1/T)\ell(T,q) &= G[F(T),q], \text{ where} \\ F(T) &= (F_i(T))_{i \in I} \ . \end{aligned} \tag{27}$$

Define the family of open sets, $B(\delta, y), \delta > 0, y >> 0 \in U$,by

$$B(\delta, y) = \{z : z \in U, G(y,y) - G(y,z) < \delta\}. \tag{28}$$

This family constitutes a base for the open sets in U. Fix $\delta > 0$ and $q' \in U$, with $q' >> 0$, and consider the sets $B = B(4\delta, q')$ and $A = B(\delta, q')$. Write $g' \equiv G(q', q')$. For $\eta > 0$ define $C(\eta)$ to be the open sphere in U with center q' and (Euclidean) diameter η. Since G is uniformly continuous on U^2, there exists $\eta > 0$ such that, for all $f, q \in U$,

$$f \in C(\eta) \Rightarrow |G(f,q) - G(q',q)| < \delta. \tag{29}$$

On the one hand,

$$\begin{aligned} f &\in C(\eta), q \in A \Rightarrow \qquad (30) \\ G(f,q) - g' &= G(f,q) - G(q',q) + G(q',q) - g' \\ &> -\delta - \delta = -2\delta, \\ &\Rightarrow G(f,q) > g' - 2\delta. \end{aligned}$$

On the other hand,

$$\begin{aligned} f &\in C(\eta), q \notin B \Rightarrow \qquad (31) \\ g' - G(f,q) &= g' - G(q',q) + G(q',q) - G(f,q) \\ &> 4\delta - \delta = 3\delta \\ &\Rightarrow G(f,q) < g' - 3\delta. \qquad (32) \end{aligned}$$

As usual, let $\tilde{}B$ denote the complement of B. Together, the last two statements imply that,

$$\begin{aligned} \text{if } F(T) &\in C(\eta), \text{ then} \\ \int_B L(T,q)d\mu(q) &\geq \int_A L(T,q)d\mu(q) > \mu(A)\exp[(g' - 2\delta)T] > 0; \\ \int_{\tilde{}B} L(T,q)d\mu(q) &< [1 - \mu(B)]\exp\left[(g' - 3\delta)T\right]; \\ \frac{\int_B L(T,q)d\mu(q)}{\int_{\tilde{}B} L(T,q)d\mu(q)} &> \rho(B)\exp\left(\delta T\right), \\ \text{where } \rho(B) &\equiv \frac{\mu(B)}{[1 - \mu(B)]}. \end{aligned}$$

Hence

$$F(T) \in C(\eta) \Rightarrow \psi(\cdot, T)(B) > \frac{\rho(B)\exp(\delta T)}{1+\rho(B)\exp(\delta T)}. \tag{33}$$

By the Strong Law of Large Numbers,

$$\lim_{T\to\infty} F(T) = q'\text{a.s.},$$

and hence there is a random time, T', such that

$$T \geq T' \Rightarrow F(T) \in C(\eta), \tag{34}$$

which in turn implies that

$$\lim_{T\to\infty} \psi(\cdot, T)(B) = 1. \tag{35}$$

To summarize, I have shown that for any $\delta > 0$, this limit holds for the set $B = B(4\delta, q')$. Hence, the limit holds for *any open set* B *containing* q', which completes the proof of the lemma.

To complete the proof of the theorem, corresponding to each of the sums (21), define

$$\begin{aligned} F_{hj}(T) &= \left(\frac{N_{hj}(T)}{N_h(T)}\right), \\ F_h(T) &= (F_{hj}(T))_{j\in I}. \end{aligned} \tag{36}$$

[Note that I show explicitly that N_{hj} and N_h depend on T.] Thus each $F_h(T)$ is a vector in the simplex U. Let q'_h denote row h of the true transition matrix, Q'. For each $\delta > 0$ and each $h \in K$, define the (open) neighborhood $B(4\delta, q'_h)$ of q'_h in U as in the proof of the lemma, and let B be the Cartesian product of these sets, i.e.,

$$B = \times_h B(4\delta, q'_h). \tag{37}$$

As one varies δ the family of sets B forms a basis for the neighborhoods of Q'. Similarly, let

$$A = \times_h B(\delta, q'_h). \tag{38}$$

For each h let $C_h(\eta, q'_h)$ be the set corresponding to $C(\eta)$ in the Lemma. By (20), for each $h \in K$ there is a random time T'_h such that

$$T \geq T'_h \Rightarrow F_h(T) \in C_h(\eta, q'_h). \tag{39}$$

Hence

$$T \geq T' \equiv \max_{h \in K} T'_h \Rightarrow \text{ for every } h \in K, \quad F_h(T) \in C_h(\eta, q'_h). \tag{40}$$

I shall now show that a calculation parallel to that in the proof of the lemma completes the proof of the theorem. Recall from (19),

$$\left(\frac{1}{T}\right) \ell_{k+T-1}(x, \theta) \equiv \left(\frac{1}{T}\right) \ln s[x(0), ..., x(k-1)] + \sum_{h \in K} \left[\frac{N_h(T)}{T}\right] G\left[F_h(T), q_h\right],$$

where $F_h(T)$ is the vector with coordinates

$$\frac{N_{hj}(T)}{N_h(T)}, \quad j \in I.$$

To simplify the notation, let

$$\begin{aligned}
C_h &= \text{ a sphere with center } q'_h \text{ such that for all } q \in U, \\
f &\in C_h \Rightarrow |G(f,q) - G(q'_h, q)| < \delta; \\
C &= \times C_h.
\end{aligned}$$

Also, let

$$F(T) = [F_h(T)]_{h \in K}.$$

On the one hand, $F(T) \in C$ and $Q \in A$ imply that, for every h,

$$\begin{aligned}
G[F_h(T), q_h] &= G[F_h(T), q_h] - G(q'_h, q_h) + G(q'_h, q_h) \\
&\geq -\delta + g'_h - \delta = g'_h - 2\delta, \text{ where} \\
g'_h &= G(q'_h, q'_h).
\end{aligned}$$

On the other hand, $F(T) \in C$ and $Q \notin B$ imply that, for every h,

$$\begin{aligned}
G[F_h(T), q_h] &= G[F_h(T), q_h] - G(q'_h, q_h) + G(q'_h, q_h) \\
&\leq \delta + g'_h - 4\delta = g'_h - 3\delta.
\end{aligned}$$

Let

$$\begin{aligned}
\theta &= (Q, s), \theta' = (Q', s'), \\
\mathbb{B} &= \{\theta : Q \in B, s \text{ invariant for } Q\}, \\
\mathbb{A} &= \{\theta : Q \in A, s \text{ invariant for } Q\}, \\
S(x, k, \theta) &= s[x(0), ..., x(k-1)].
\end{aligned}$$

Then $F(T) \in C$ and $\theta \in \mathbb{A}$ imply that

$$\left(\frac{1}{T}\right) \ell_{k+T-1}(x,\theta) \geq \left(\frac{1}{T}\right) \ln S(x,k,\theta) + \sum_{h \in K} \left[\frac{N_h(T)}{T}\right] g'_h - 2\delta,$$

and $F(T) \in C$ and $\theta \notin \mathbb{B}$ imply that

$$\left(\frac{1}{T}\right) \ell_{k+T-1}(x,\theta) \leq \left(\frac{1}{T}\right) \ln S(x,k,\theta) + \sum_{h \in K} \left[\frac{N_h(T)}{T}\right] g'_h - 3\delta.$$

Since Q' is irreducible, for every h,

$$\begin{aligned} \lim_{T\to\infty} \frac{N_h(T)}{T} &= s'_h, \\ \lim_{T\to\infty} F(T) &= Q'. \end{aligned}$$

Hence, eventually (a.s.)

$$\left| \sum_{h \in K} \left(\frac{N_h(T)}{T} - s'_h \right) g'_h \right| \leq \frac{\delta}{3},$$

and so, eventually,

$$\begin{aligned} \theta \in \mathbb{A} &\Rightarrow \left(\frac{1}{T}\right) \ell_{k+T-1}(X,\theta) \geq \left(\frac{1}{T}\right) \ln S(X,k,\theta) + \sum_{h \in K} s'_h g'_h - \frac{7\delta}{3}, \\ \theta \notin \mathbb{B} &\Rightarrow \left(\frac{1}{T}\right) \ell_{k+T-1}(X,\theta) \leq \left(\frac{1}{T}\right) \ln S(X,k,\theta) + \sum_{h \in K} s'_h g'_h - \frac{8\delta}{3}. \end{aligned}$$

Note that

$$\frac{\psi(T)(\mathbb{B})}{\psi(T)(\tilde{\ }\mathbb{B})} \geq \frac{\psi(T)(\mathbb{A})}{\psi(T)(\tilde{\ }\mathbb{B})}.$$

Hence, eventually,

$$\begin{aligned}
\frac{\psi(T)(\mathbb{B})}{\psi(T)(\tilde{}\mathbb{B})} &\geq \frac{\int_{\mathbb{A}} S(X,k,\theta)\exp\{T\left[\sum_{h\in K} s'_h g'_h - \frac{7\delta}{3}\right]\}d\mu(\theta)}{\int_{\tilde{}\mathbb{B}} S(X,k,\theta)\exp\{T\left[\sum_{h\in K} s'_h g'_h - \frac{8\delta}{3}\right]\}d\mu(\theta)} \\
&= \left[\frac{\int_{\mathbb{A}} S(X,k,\theta)d\mu(\theta)}{\int_{\tilde{}\mathbb{B}} S(X,k,\theta)d\mu(\theta)}\right]\exp\{\frac{T\delta}{3}\} \\
&\to \infty \text{ as } T \to \infty,
\end{aligned}$$

and so

$$\pi \lim_{T\to\infty} \psi(T)(\mathbb{B}) = 1,$$

which completes the proof of the theorem.

Suppose that the true $p' \in S$ is stationary, but the DM's prior assigns probability one to $M(k)$, and p' is not in $M(k)$. What will happen to the DM's posteriors? The following corollary, which is stated without proof, provides an answer this question.

Corollary 8 *Suppose that*

(1) p' *determines the stochastic process* $\{X(t)\}$*;*

(2) $\mu[M(k)] = 1$*;*

(3) $Q(k,p')$*, the conditional k'th-order transition probability matrix implied by* p'*, is irreducible, with stationary probability vector* $s(k,p')$*; and define*

(4) $\theta(k,p') = [Q(k,p'), s(k,p')]$;

then the sequence of posteriors converges to the Dirac measure $D[\theta(k,p')]$ *a.s. with respect to the measure* p'*.*

3.2 Bayesian Analysis in M

In this subsection I show that consistent Bayesian analysis is possible for a large class of priors on the entire space M. For each k, let μ_k be a prior such that $\mu_k[M(k)] = 1$, and that has full support on $M(k)$. Let $\{\alpha_k\}$ be a sequence of strictly positive numbers whose sum is unity, and define

$$\mu = \sum_{k=0}^{\infty} \alpha_k \mu_k.$$

From (15), the posterior distribution is

$$\psi(x,T)(A) \equiv \frac{\sum_{k=0}^{\infty} \alpha_k \int_A L_T(x,\theta) d\mu_k(\theta)}{\sum_{k=0}^{\infty} \alpha_k \int_M L_T(x,\theta) d\mu_k(\theta}.$$

Recall that θ'denotes the value of θ that generates X_∞.

Theorem 9 *Suppose that $\theta' \in M(k')$ is irreducible; then μ is consistent for θ'.*

Proof. Let $A \subset M$ be open and containing θ', such that A is disjoint from $M(k)$ for $k < k'$. Then we want to show that

$$\lim_{T\to\infty} \psi(X_\infty, T)(A) = 1 \text{ a.s.}$$

For every k, the posterior probability measure corresponding to the prior μ_k is

$$\psi_k(X_\infty, T)(A) \equiv \frac{\int_A L_T(X_\infty,\theta) d\mu_k(\theta)}{\int_M L_T(X_\infty,\theta) d\mu_k(\theta}.$$

By Theorem 6, for each $k \geq k'$, this posterior probability converges a.s. to one as T increases without bound. Hence there is a (measurable) subset C of X with

θ'-probability one such that for every $x \in C$,

$$\text{for all } k,\ \lim_{T\to\infty} \psi_k(x,T)(A) = 1.$$

Fix $x \in C$, and for the purposes of this proof define

$$\begin{aligned} Y_T(k) &= \int_A L_T(x,\theta)d\mu_k(\theta), \\ Z_T(k) &= \int_M L_T(x,\theta)d\mu_k(\theta); \end{aligned}$$

then

$$\psi(x,T)(A) = \frac{\sum_{k=0}^{\infty} \alpha_k Y_T(k)}{\sum_{k=0}^{\infty} \alpha_k Z_T(k)}.$$

Note that $0 \le Y_T(k) \le Z_T(k) \le 1$, and $Y_T(k) = 0$ for $k < k'$. Define

$$\begin{aligned} \beta_k &= \frac{\alpha_k}{\sum_{m \ge k'} \alpha_m},\ \text{ for } k \ge k', \\ &= 0,\ \text{ for } k < k'; \end{aligned}$$

then

$$\psi(x,T)(A) = \frac{\sum_k \beta_k Y_T(k)}{\sum_k \beta_k Z_T(k)}.$$

Define

$$f_T(k) = \frac{Z_T(k)\left[1 - \frac{Y_T(k)}{Z_T(k)}\right]}{\sum_m \beta_m Z_T(m)},\quad \text{for } k \ge k'.$$

Observe that

$$0 \le f_T(k) \le 1 - \frac{Y_T(k)}{Z_T(k)} \le 1.$$

By Theorem 6,

$$\lim_{T\to\infty} f_T(k) = 0, \text{ for } k \ge k'.$$

Think of $\{\beta_k\}$ as a finite measure on the integers $\geq k'$; then by the Dominated Convergence Theorem (or use an elementary argument),

$$\lim_{T\to\infty} \sum_k \beta_k f_T(k) = 0.$$

From the definition of $f_T(k)$,

$$\begin{aligned}
\sum_k \beta_k f_T(k) &= \frac{\sum_k \beta_k Z_T(k)\left[1 - \frac{Y_T(k)}{Z_T(k)}\right]}{\sum_m \beta_m Z_T(m)} \\
&= \frac{\sum_k \beta_k \left[Z_T(k) - Y_T(k)\right]}{\sum_m \beta_m Z_T(m)} \\
&= 1 - \frac{\sum_k \beta_k Y_T(k)}{\sum_m \beta_m Z_T(m)} \\
&= 1 - \psi(x,T)(A).
\end{aligned}$$

Hence

$$\lim_{T\to\infty} [1 - \psi(x,T)(A)] = 0,$$

which completes the proof.

It is not known to me whether the last theorem can be generalized to cover, S, the class of all stationary processes in this setting. (For related work, see Sims, 1971.)

4 A "Bayesian" Analysis of Model Revision

Suppose now that the parameter k is not known. If an upper bound on k, were known, say k'', then a Bayesian could estimate θ consistently by taking a suitable prior on $M(k'')$. However, if an upper bound is not known, then the Bayesian must take a prior on M, which is infinite dimensional. As noted in the Introduction, it might be difficult for an individual to formulate beliefs about such a large space that are sufficiently precise to yield a meaningful prior. In this section I outline an alternative approach, which is a process of *model revision* suggested by considerations of *bounded rationality.* This process yields estimates of θ in M that are "consistent" in a sense to be defined below.

In decision-making in the economic sphere, the decision-maker (henceforth DM) usually starts with a model of behavior of the relevant agents (consumers, firms, traders), but the model has unknown parameters. Suppose that the DM's model implies that the process, $\{X(t)\}$, is a stationary k'th-order Markov process, i.e., the DM assumes that the probability measure p in $P(X)$ is in $M(k)$, for a particular value of k, say k_1. Let p' denote the *true* probability measure, and assume that p' is in M but not necessarily in $M(k_1)$. Assume, also, that p' is irreducible. Finally, *assume that the DM's prior probability measure on $M(k_1)$, say μ^1, has full support.*

According to Corollary 8, in this situation the DM's posterior distributions will converge to a distribution concentrated on a point in $M(k_1)$, namely

$$p'(k_1) \equiv D[Q(k_1, p'), s(k_1, p')].$$

If p' is in $M(k_1)$, then $p'(k_1) = p'$. However, if p' is not in $M(k_1)$, then $p'(k_1)$ will not equal p', although it will be consistent with p' on cylinder sets in X of length $(k+1)$. If the DM uses only Bayes's Theorem to make inferences about p, then he will only calculate statistics about cylinder set of length $(k_1 + 1)$, and *he will never learn that the true probability measure is not in* $M(k_1)$.

Suppose now that the DM starts to consider a different model. There are various reasons that this might happen. For example, new developments in economics or other social sciences might lead him to suspect that his original model is not adequate. The DM might also engage in some "data mining" and simply explore the statistics of cylinder sets of length greater than (k_1+1). Suppose that the new model implies that the process p is in $M(k_2)$ with $k_2 > k_1$. This last inequality is more of a convention than an assumption. If k_2 were less than or equal to k_1 then the DM would suffer no loss by continuing with the hypothesis that $p \in M(k_1)$, except possibly a loss of efficiency of estimation from using a larger parameter space than necessary.

We can now imagine that this procedure is repeated indefinitely. (I shall describe the procedure and its implications more formally below.) Roughly speaking, it will result in a sequence $\{ \mu^n\}$ of priors with larger and larger supports, $M(k_n)$. To each prior will correspond a (stochastic) process of posteriors, $\psi^n(X_\infty, T)$, with the properties described above. Recall that $p'$is the true probability measure governing the stochastic process $\{X(t)\}$, and supppose that p' is in $M(k')$. Again, by Corollary 8, for each model n ,

$$\lim_{T\to\infty} \psi^n(X_\infty, T) = D[p'(k_n)]. \tag{41}$$

Furthermore,

$$\lim_{T\to\infty} \psi^n(X_\infty, T) = p' \text{ if } k_n \geq k'.$$

However, these last equations are not in themselves enough to characterize the DM's process of inference. For this, it will be necessary to specify what posterior, say $\Psi(T)$, the DM will be using at each date T. . Let $n(T)$ denote the model that the DM is using at date T; then

$$\Psi(T) = \psi^{n(T)}(X_\infty, T). \tag{42}$$

The goal is to choose successive models and priors so that, in some sense, the posteriors $\Psi(T)$ eventually get "close" to p'. To achieve this, the DM needs to have enough evidence about any given model. The larger the model, i.e., the larger th*e space* $M(k_n)$, the slower will be the convergence of the corresponding posteriors. *Hence the DM must not change models too quickly.*

I shall now describe the process of model revision more precisely. Let $\{k_n\}$ and $\{t_n\}$ be strictly increasing (deterministic) sequences of integers, and suppose that the DM switches to model $M(k_n)$ at date t_n. For the moment, think of the sequence of dates as deterministic. Corresponding to each n is a prior, μ^n, on $M(k_n)$ and to each T such that $t_n \leq T < t_{n+1}$, a posterior, $\psi^n(X_\infty, T)$. .

I now present the main result of this section. To simplify the notation, and without any essential loss of generality, assume that each model revision increases k by 1, so that $k_n = n$. In fact, I shall suppress the symbol n altogether, so that the DM switches to model $M(k)$ at date $t(k)$. Accordingly, at a date T such that $t(k) \leq T < t(k+1)$, the DM's prior on $M(k)$ is denoted by μ^k and the corresponding posterior by $\psi^k(T)(x)$ for the particular realization x of the stochastic process X_∞ (see Section 3). As noted at the beginning of Section 3, the probability measure μ^k on $M(k)$ induces a probability measure, say ν^k, on $Y(k) = M(k) \times X$.

For $\delta > 0$ and $\theta \in M(k)$ let $V^k(\theta, \delta)$ denote the interior of the Euclidean ball in $M(k)$ with center θ and radius δ, and for any $x \in X$ define

$$m_T^k(x, \theta, \delta) = \psi^k(x, T)(x)V^k(\theta, \delta). \tag{43}$$

Note that $m_T^k(X_\infty, \theta, \delta)$ is a random variable on the probability space $[Y(k), \nu^k]$.

Finally, let $M'(k)$ denote the set of irreducible θs in $M(k)$, i.e.,

$$M'(k) = \{\theta \in M(k) | \theta \text{ is irreducible}\},$$

and let

$$Y'(k) = M'(k) \times X.$$

Theorem 10 *Suppose that, for each k, $\mu^k [M'(k) = 1]$. Let $\{\delta_k\}, \{\epsilon_k\}$ be any two sequences of strictly positive numbers converging to 0; then there exist a sequence of times, $\{t(k)\}$, and a sequence of sets, $\{A(k)\}$, such that*

$$A(k) \subset Y'(k), \tag{44}$$

$$\nu^k[A(k)] \geq 1 - \epsilon_k, \text{ and} \tag{45}$$

$$T \geq t(k) \Rightarrow m_T^k(x, \theta, \delta_k) \geq 1 - \epsilon_k \text{ for all } (x, \theta) \in A(k).$$

Proof. The proof is a straightforward application of Egoroff's Theorem (e.g., Halmos (1950), p. 88). To begin the proof, fix k and δ. By Theorem 6, *conditional on* $\theta \in M'(k)$,

$$\lim_{T\to\infty} m_T^k(x,\theta,\delta) = 1, \text{ a.s., given } \theta. \tag{46}$$

Hence the same statement is true *unconditionally* on the space $Y'(k)$. By Egoroff's Theorem, this convergence is "almost uniform," i.e., for every $\epsilon > 0$ there exists a set $A'(\delta,\epsilon) \subset Y'(k)$ such that

$$\begin{aligned} \nu^k[A'(\delta,\epsilon)] &\geq 1-\epsilon, \\ m_T^k(x,\theta,\delta) &\rightarrow 1 \text{ uniformly in } A'(\delta,\epsilon). \end{aligned}$$

Hence there exists $T(\delta,\epsilon)$ such that

$$T \geq T(\delta,\epsilon) \Rightarrow m_T^k(x,\theta,\delta) \geq 1-\epsilon \text{ for all } (x,\theta) \in A'(\delta,\epsilon).$$

To complete the proof, take

$$\begin{aligned} t(k) &= T(\delta_k,\epsilon_k), \\ A(k) &= A'(\delta_k,\epsilon_k). \end{aligned} \tag{47}$$

The conclusion of the preceding theorem is different in a subtle but important way from that of Theorem 6. It states that *the DM is increasingly confident that the current posterior is increasingly concentrated on the true* θ.

The analysis of two interesting questions lies beyond the scope of this paper. First, it would be useful to know something about how fast $t(k)$ increases with k, and how the sequence $\{t(k)\}$ depends on δ and ϵ. Note that the dimension of $M(k)$ increases geometrically with k. Second, in the light of Proposition 5, one might conjecture that an analogue of the preceding Theorem would be true for the entire set of stationary processes.

5 References

Diaconis, Persi, and David Freedman, "On the Consistency of Bayes Estimates," Annals of Statistics, 14, no. 1 (1986), 1-26. [Review paper]

Doob, J. L., *Stochastic Processes,* John Wiley, New York, 1953.

Freedman, David, "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case, I, "*Annals of Mathematical Statistics*, 34 (1963), 1386-1403.

Freedman, David, "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case, II," *Annals of Mathematical Statistics*, 36 (1965), 454-456.

Halmos, *Measure Theory*, D. Van Nostrand, New York, 1950.

Parthasarathy, K. R., *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

Radner, Roy, "Bounded Rationality, Indeterminacy, and the Theory of the Firm," *Economic J.*, 106, no. 438 (Sept. 1996), 1360-1373.

Radner, Roy, "Costly and Bounded Rationality in Individual and Team Decision-Making," *Industrial and Corporate Change*, v.9, no. 4 (2000), 623-658.

Savage, Leonard J., *The Foundations of Statistics*, John Wiley, New York, 1954 (2nd ed.), Dover, 1972.

Sims, Christopher, "Distributed Lag Estimation when the Parameter Space is Explicitly Infinite Dimensional," *Annals of Mathematical Statistics*, 42, no. 5 (1971), 1622-1636.