Cowles Foundation Paper 917

Econometrica, Vol. 64, No. 2 (March, 1996), 381–412

AN ASYMPTOTIC THEORY OF BAYESIAN INFERENCE FOR TIME SERIES

BY PETER C. B. PHILLIPS AND WERNER PLOBERGER[1]

This paper develops an asymptotic theory of Bayesian inference for time series. A limiting representation of the Bayesian data density is obtained and shown to be of the same general exponential form for a wide class of likelihoods and prior distributions. Continuous time and discrete time cases are studied. In discrete time, an embedding theorem is given which shows how to embed the exponential density in a continuous time process. From the embedding we obtain a large sample approximation to the model of the data that corresponds to the exponential density. This has the form of discrete observations drawn from a nonlinear stochastic differential equation driven by Brownian motion. No assumptions concerning stationarity or rates of convergence are required in the asymptotics. Some implications for statistical testing are explored and we suggest tests that are based on likelihood ratios (or Bayes factors) of the exponential densities for discriminating between models.

KEYWORDS: Autoregression, Bayesian data measure, data density process, Doléans exponential, exponential data density, likelihood, martingale, posterior process, prior density, quadratic variation process, stochastic differential equation, unit root.

1. INTRODUCTION

THE BAYESIAN APPROACH TO MODELING and inference in time series econometrics have become increasingly popular in recent years. Time series applications raise concerns that deserve special attention, like the nature of prior information in time series models, the treatment of initial conditions, and nonstationarity. These concerns are the subject of two recent themed issues of the *Journal of Applied Econometrics* (1991) and *Econometric Theory* (1994). The focus of attention in the present paper is not the aforementioned concerns *per se*, but the development of a general asymptotic theory of Bayesian inference for time series. As a complement to the literature on formulating "uninformative priors" for time series, this paper is aimed at obtaining asymptotic results whereby the prior is dominated by the data.

Our starting point is to obtain an asymptotic representation of the distribution of the data that is implied by the use of Bayesian methods, i.e. the so-called Bayesian data density (which commonly figures as the proportionality factor in

[1] The authors thank John Hartigan and Bent Sorensen for comments on the first version and a co-editor and three referees for comments on three previous versions of this paper which were circulated in March 1991, October 1992, and February 1994. The first version of the paper was written while Werner Ploberger was a visitor at the Cowles Foundation during 1990–1991 and was entitled "Time Series Modelling with a Bayesian Frame of Reference: Concepts, Illustrations, and Asymptotics." Phillips thanks the NSF for research support under Grant Nos. SES 8821180 and SES 9122142. Ploberger thanks the Fonds zur Förderung der wissenschaftlichen Forschung for supporting his stay at Yale with Schrödingerstipendium Nr. J0469-TEC. The authors' thanks also go to Glena Ames for her skill and effort in keyboarding the manuscript.

Bayesian posterior distribution calculations). We show that the Bayesian data density is asymptotically of the same general exponential form for a wide class of likelihood functions and prior densities. This includes nonstationary as well as stationary systems, and no specific rates of convergence are required for our asymptotic theory to hold. The exponential data density has a simple form and depends only on the likelihood score process and its conditional variance. When we condition on a minimal information time (which is useful when making comparisons between models), the exponential density is also independent of the prior and it can be treated as a proper probability density (in the sense that its mass is unity) even when the prior is improper by making a time change in the process. The exponential density can be used to conduct Bayesian likelihood ratio tests in much the same way as Bayes factors are presently used (e.g., see Berger (1985, Section 4.3)). These tests are useful in evaluating competing statistical hypotheses about the model generating the data and, more generally, as a model selection device. The latter procedure is closely related to the BIC order selection device of Schwarz (1978) and the posterior odds criteria discussed by Leamer (1978) and Zellner (1978). Our approach allows explicitly for nonstationary data and for improper priors and yields convenient and robust characterizations of Bayes factors for use in model selection. We illustrate its use in the context of unit root tests.

Some use is made of continuous martingales and their stochastic calculus in our development and the reader is referred to Ikeda and Watanabe (1989) and Protter (1990) for the background theory. The following notational conventions are employed. V_t is used to represent a continuous L_2 (i.e. square integrable) martingale, or local martingale, and the square bracket $[V]_t = [V, V]_t$ denotes its quadratic variation process. Similar notation is employed in the case of a discrete time martingale V_n, and in this case we use $\langle V \rangle_n = \langle V, V \rangle_n$ to denote the conditional quadratic variation process. A_t (respectively, B_n) is often a shorthand notation for quadratic variation process (respectively, conditional quadratic variation). W_t denotes standard Brownian motion which is signified by the symbolism "BM(1)." The symbol " $\equiv$ " signifies equivalence or equivalence in distribution, RN derivative is short for Radon-Nikodym derivative and " $\ll$ " denotes the absolute continuity operator. We use $\lambda_{\min}(A)$ for the smallest latent value of the square matrix A, $\|A\| = \{\mathrm{tr}(A'A)\}^{1/2}$ for Euclidean distance, and $|A|$ for $\det(A)$.

2. APPROXIMATING THE BAYESIAN DATA MEASURE BY AN EXPONENTIAL BAYES MEASURE

Our analysis starts with continuous time processes and we later (in Section 3.1) show how the corresponding theory in discrete time can be embodied in that of the continuous time case by suitable embedding techniques. One of our objectives is to provide a large sample Bayesian criterion for a particular model selection problem that will be extended in Section 4. The problem is to choose between a scalar parameterized family of densities and a prespecified density within this family. This stylized problem is a useful starting point in the study of

the general problem of model selection, and posing the data evolution in continuous time leads to a convenient and elegant solution. In particular: (i) we can use the theory of continuous square integrable martingales and exponential martingales to achieve some simple but general results; (ii) we do not need to distinguish stationary, nonstationary, and explosive cases (as is usually the case in a discrete time analysis); and (iii) determination of the model for the data that is implied by the use of Bayes rule is quite simple for continuous processes.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(\mathcal{F}_t)_{t \geq 0}$ be an increasing family of right continuous sub σ-fields of $\mathcal{F}$. Assume that we have given a parameterized family of probability measures P_t^θ on the sequence of filtered spaces $(\Omega, \mathcal{F}_t)$. In this section we will consider the case of a scalar $\theta \in \mathbb{R}$ and we will assume that $P_t^\theta \ll \nu$, some σ-finite measure on $(\Omega, \mathcal{F}_t)$. We can think of P_t^θ as the probability measure of a continuous time random process $(Y_s)_{s \leq t}$ on $(\Omega, \mathcal{F}_t)$. If $\pi(\theta)$ is a prior density on θ, then the mixture

$$\mathcal{P}_t = \int_{\mathbb{R}} \pi(\theta) P_t^\theta \, d\theta \tag{1}$$

is the Bayesian data measure, i.e. the (probability) measure of the data $Y^t = (Y_s)_{s \leq t}$, with the corresponding data density $d\mathcal{P}_t/d\nu = \int_{\mathbb{R}} \pi(\theta)(dP_t^\theta/d\nu)\, d\theta$. We put the word probability in parentheses in the last sentence because $\mathcal{P}_t$ may not be a proper probability measure in the sense that it has unit mass. For instance, if $\pi(\theta)$ is improper then $\mathcal{P}_t$ is σ-finite with mass $\mathcal{P}_t(\Omega) = \int_{\mathbb{R}} \pi(\theta) P_t^\theta(\Omega)\, d\theta = \int_{\mathbb{R}} \pi(\theta)\, d\theta = \infty$. We accommodate this possibility in what follows. Let θ^0 be the "true value" of θ and set $P_t^0 = P_t^{\theta^0}$. We write the likelihood function in terms of the density $L_t(\theta) = dP_t^\theta/dP_t^0$ and then the density of $\mathcal{P}_t$ with respect to P_t^0 is

$$d\mathcal{P}_t/dP_t^0 = \int_{\mathbb{R}} \pi(\theta)(dP_t^\theta/dP_0^t)\, d\theta = \int_{\mathbb{R}} \pi(\theta) L_t(\theta)\, d\theta. \tag{2}$$

Our object is now to show that $\mathcal{P}_t$ can be approximated asymptotically by a much simpler measure, which we denote by Q_t, and to find its general form.

2.1. THEOREM: *Assume the following conditions hold:*

(C1) $l_t(\theta) = \ln(L_t(t))$ *is twice continuously differentiable with derivatives* $l_t^{(1)}(\theta)$ *and* $l_t^{(2)}(\theta)$.

(C2) *Under* P_t^θ, $l_t^{(1)}(\theta)$ *is a continuous local martingale with quadratic variation process* $A_t(\theta)$ *and* $A_t(\theta) \to \infty$ *a.s.* (P^θ) *as* $t \to \infty$. *Let* $A_t = A_t(\theta^0)$.

(C3) $(l_t^{(2)}(\theta) + A_t(\theta))/A_t(\theta) \to 0$ *a.s.* (P^θ) *as* $t \to \infty$.

(C4) *There exist continuous functions* $w_t(\theta, \theta')$ *such that* $w_t(\theta, \theta) = 0$ *and such that for all* θ, θ' *in some neighborhood* $N_\delta(\theta^0) = \{\theta : |\theta - \theta^0| < \delta\}$ *of* θ^0 *we have*

$$\{l_t^{(2)}(\theta) - l_t^{(2)}(\theta')\}/A_t \leq w_t(\theta, \theta') \text{ a.s. } (P^0)$$

for each t, *and* $w_t(\theta, \theta') \to w_\infty(\theta, \theta')$ *a.s.* (P^0) *uniformly for* $\theta, \theta' \in N_\delta(\theta^0)$ *and* $w_\infty(\theta, \theta) = 0$.

(C5) *The maximum likelihood estimate* $\hat{\theta}_t$ *for* θ^0 *is consistent, i.e.* $\hat{\theta}_t \to \theta^0$ *a.s.* (P^0).

(C6) *For any* $\delta > 0$ *and* $\omega_\delta = \{\theta : |\theta - \theta^0| \geq \delta\}$ *we have*

$$A_t^{1/2} \int_{\omega_\delta} \pi(\theta)(dP_t^\theta / dP_t^0)\, d\theta \to 0 \ a.s.\ (P^0).$$

(C7) *The prior density* $\pi(\theta)$ *is continuous at* θ^0 *with* $\pi_0 = \pi(\theta^0) > 0$. *Then*

$$\frac{d\mathscr{P}_t}{dP_t^0} \Big/ \frac{dQ_t}{dP_t^0} \to 1 \ a.s.\ (P^0), \tag{3}$$

where Q_t *is the measure defined by the following RN derivative with respect to* P_t^0:

$$\frac{dQ_t}{dP_t^0} = c_0 \frac{\exp\{(1/2)V_t^2 A_t^{-1}\}}{A_t^{1/2}}, \tag{4}$$

where $V_t = l_t^{(1)}(\theta^0)$, *and* $c_0 = (2\pi)^{1/2}\pi_0$. *The derivative* (4) *may also be written in the following asymptotically equivalent forms*:

$$dQ_t/dP_t^0 = c_0 \exp\left\{(1/2)\left(\hat{\theta}_t - \theta^0\right)^2 A_t\right\} \Big/ A_t^{1/2} \tag{5}$$

and

$$dQ_t/dP_t^0 = c_0 \exp\left\{l_t\left(\hat{\theta}_t\right)\right\} \Big/ A_t^{1/2}. \tag{6}$$

2.2. REMARKS ON (C1)–(C7): (i) Condition (C1) is standard in the asymptotic theory of regular estimators. So is the first part of (C2)—in the usual maximum likelihood theory (e.g., Hall and Heyde (1980, p. 157)) $l_t^{(1)}(\theta)$ is a continuous L_2 martingale under P_t^θ. The conditional variance process of $l_t^{(1)}(\theta)$ (the time clock of the martingale) is the quadratic variation or square bracket process $A_t(\theta) = [l^{(1)}(\theta)]_t$. The requirement that $A_t(\theta) \to \infty$ a.s. (P^θ) ensures that there is eventually an infinite amount of information about the process in the likelihood function. It corresponds to the usual persistent excitation condition in regression theory.

(ii) Condition (C3) says that $l_t^{(2)}(\theta) + A_t(\theta)$ must be small relative to $A_t(\theta)$ as $t \to \infty$. This is a version of the usual requirement that $J_n/I_n \to -1$ as $n \to \infty$ in the theory of maximum likelihood where I_n is the conditional variance of the score and J_n is the Hessian of the likelihood (see, for instance, Hall-Heyde (1980, p. 160)). In the present case, note that $l_t^{(2)}(\theta) + (l_t^{(1)}(\theta))^2$ is a local martingale (as in the standard ML theory) and, moreover, since $l_t^{(1)}(\theta)$ is a local P^θ-martingale, so also is $(l_1^{(1)}(\theta))^2 - A_t(\theta)$. Hence, we would expect the sum of these local martingales,

$$l_t^2(\theta) + A_t(\theta) = \left\{l_t^2(\theta) + (l_t^{(1)}(\theta))^2\right\} + \left\{A_t(\theta) - (l_t^{(1)}(\theta))^2\right\}$$

to be small relative to the quadratic variation process $A_t(\theta)$, thereby giving intuitive support to (C3).

(iii) Condition (C4) is a smoothness condition. It requires, in effect, that relative differences in $l_t^{(2)}(\theta)$ and $l_t^{(2)}(\theta')$ be bounded above by an equicontinuous family of functions $\omega_t(\theta, \theta')$ in some neighborhood of θ^0 with the property that when $\theta = \theta'$ the limit function $\omega_\infty(\theta, \theta) = 0$.

(iv) Condition (C5) is standard. It could be replaced by an explicit condition on the behavior of the likelihood ratio dP_t^θ / dP_t^0 as $t \to \infty$ in closed sets like $\omega_\delta = \{\theta : |\theta - \theta^0| \geq \delta\}$ that do not contain θ^0. For instance, one commonly occurring condition (e.g. Walker (1969, p. 83) and Hall-Heyde (1980, p. 158)) would, in the present case, take the form that for every $\delta > 0$ there is a $k(\delta) > 0$ such that

$$P\left(\sup_{\theta \in \omega_\delta} \frac{dP_t^\theta}{dP_t^0} < \exp\{-A_t(\theta^0)k(\delta)\} \right) \to 1.$$

A somewhat stronger version of this condition is that for $\delta > 0$ there exists a $k(\delta)$ such that

$$\text{(C5')} \qquad \exp\{A_t(\theta^0)k(\delta)\} \sup_{\theta \in \omega_\delta} \frac{dP_t^\theta}{dP_t^0} < 1 \text{ a.s. } (P^0)$$

as $t \to \infty$. Then, if the prior density $\pi(\theta)$ were proper, we would have

$$\begin{aligned} \text{(C6')} \qquad & A_t(\theta^0)^{1/2} \int_{\omega_\delta} \pi(\theta)(dP_t^\theta / dP_t^0)\, d\theta \\ & \leq A_t(\theta^0)^{1/2} \sup_{\theta \in \omega_\delta} (dP_t^\theta / dP_t^0) \int_{\omega_\delta} \pi(\theta)\, d\theta \\ & \leq A_t(\theta^0)^{1/2} \exp\{-A_t(\theta^0)k(\delta)\} \to 0 \text{ a.s. } (P^0), \text{ as } t \to \infty \end{aligned}$$

in view of (C5'). Result (C6') is the natural alternative to condition (C6) when the prior density $\pi(\theta)$ is proper. As it stands, (C6) simply requires that the average density dP_t^θ / dP_t^0 over a closed set like ω_δ that does not contain θ^0 be small relative to $A_t^{1/2}$ as $t \to \infty$. When $\pi(\theta)$ is proper, (C6') shows that the average density, which in this case is $\int_{\omega_\delta} \pi(\theta)(dP_t^\theta / dP_t^0)\, d\theta$, is exponentially small in $A_t(\theta^0)$ as $t \to \infty$. The explicit condition (C6) does not therefore seem to be overly strong and allows us the extra convenience of working with improper as well as proper prior densities.

(v) If the prior $\pi(\theta)$ is uniform at the constant level $\pi_0 = (2\pi)^{-1/2}$, then $c_0 = 1$ and we have the simpler forms

(4', 5', 6')

$$\begin{aligned} \frac{dQ_t}{dP_t^0} &= \frac{\exp\{(1/2)V_t^2 A_t^{-1}\}}{A_t^{1/2}} \\ &= \frac{\exp\left\{(1/2)\left(\hat{\theta}_t - \theta^0\right)^2 A_t\right\}}{A_t^{1/2}} = \frac{\exp\left\{l_t\left(\hat{\theta}_t\right)\right\}}{A_t^{1/2}} \end{aligned}$$

in place of (4), (5), and (6).

(vi) If the almost sure convergence conditions in (C2), (C3), (C5), and (C6) are replaced by convergence in probability, and if the inequality in (C4) holds with probability tending to one as $t \to \infty$, then the main result, (3), of the theorem holds in probability rather than almost surely. This weakening of the conditions may make them easier to verify in specific cases. A similar remark applies to Theorems 3.1 and 4.1 below.

2.3. REMARKS ON THEOREM 2.1: (i) We can write result (3) in the form $d\mathscr{P}_t/dQ_t \to 1$ a.s. (P^0), which tells us that the measure Q_t is identical to the Bayes data measure $\mathscr{P}_t$ as $t \to \infty$. The advantage of Q_t is that it is a data measure of the same simple form for a broad range of likelihoods and prior densities. It depends only on the score process V_t, the quadratic variation process A_t, and the value π_0 of the prior at θ^0. The density ratio (4) that defines Q_t can be written as

$$\mathscr{R}(t) = dQ_t/dP_t^0 = c_0 \exp\{K(t)\}, \quad \text{with}$$

$$K(t) = V(t)^2/2A(t) - (1/2)\ln(A(t)).$$

Using a stopping time sequence $(\tau_a)_{a \geq 0}$, such as that given in (9) below, we construct the time changed density process

$$\mathscr{R}_a = \mathscr{R}(\tau_a)/\mathscr{R}(\tau_0) = \exp\{K(\tau_a) - K(\tau_0)\} = \exp\{G_a - (1/2)[G]_\alpha\}, \tag{7}$$

where $G_a = G(\tau_a) = \int_{\tau_0}^{\tau_a} (V(t)/A(t))\, dV(t)$—cf. equations (A19)–(A21) in the Appendix. The process $\mathscr{R}_a$ given in the final equality of (14) is called a Doléans exponential (cf. Meyer (1989, p. 148 in the appendix to Emery (1989))) and is an exponential martingale when the sequence τ_a is chosen as in (9). This exponential is especially interesting in the statistical theory of stochastic processes because it is known to represent the limit of the likelihood function for stochastic processes in very general situations (see Strasser (1986, Theorem 1.15)). Moreover, as shown in the proof of Theorem 2.4 below, we here have that $E(\mathscr{R}_a) = 1$ and $\mathscr{R}_a$ therefore represents a proper probability density. For these reasons we refer to $\mathscr{R}_a$ as an exponential density and the underlying measure Q_t from which it is derived as an exponential measure.

(ii) Using (4) we have

$$2\ln(dQ_t/dP_t^0) = V_t^2 A_t^{-1} - \ln(A_t) + 2\ln(c_0),$$

the first term of which is a quadratic form in the score $V_t = l_t^{(1)}(\theta^0)$ that corresponds to the classical score statistic. The statistic dQ_t/dP_t^0, can therefore be interpreted as a form of penalized score test in which the size of the penalty (for estimating θ) is determined by the quadratic variation A_t. Note that (4) can, in fact, be computed under the null hypothesis $\theta = \theta^0$, just as the classical score or LM test, simply by using the value of θ under the null.

(iii) Formulae (5) and (6) give alternative (asymptotically equivalent) expressions for dQ_t/dP_t^0. Of particular interest is (6) because we can rewrite this in the following form (up to a constant):

$$\ln(dQ_t/dP_t^0) = l_t\left(\hat{\theta}_t\right) - (1/2)\ln A_t, \tag{8}$$

which is related to the BIC criterion of Schwarz (1978). Here $l_t(\hat{\theta}_t)$ is the maximized value of the log-likelihood (ratio) function (compare the $\log(\hat{\sigma}^2)$ expression, where $\hat{\sigma}^2$ is the residual variance, that appears in the common regression form of the BIC criterion) and $\ln(A_t)$ is the penalty term for including θ as a free parameter (i.e. free rather than set equal to the fixed value $\theta = \theta^0$, as in the competing model for which P_t^0 is the probability measure).

Formula (8) can be used explicitly as a model selection criterion to choose between a model based on the scalar θ parameterized family of distributions (for which Q_t is an asymptotic characterization of the data measure) and the prespecified measure P_t^0 within this family.

As an illustration of the use of (8) as a model selection device, consider the simple case of a Wiener process with drift, i.e. $X(t) = \theta t + W(t)$ where $W(t)$ is a standard Wiener process. We will choose the model without drift if $dQ_t/dP_t^0 < 1$, i.e. if P_t^0 has the greater likelihood. The log likelihood is

$$l_t(\theta) = \ln(dP_t^\theta/dP_t^0) = \theta\int_0^t dX - (1/2)\theta^2\int_0^t ds = \theta X(t) - (1/2)\theta^2 t.$$

The MLE is $\hat{\theta}_t = X(t)/t$, or $W(t)/t$ when P_t^0 holds. Thus, under P_t^0 the criterion (8) is simply $(1/2t)W(t)^2 + (1/2)\ln(t)$. By the law of the iterated logarithm for Brownian motion we have $W(t)^2/t = O(\ln\ln(t))$ a.s., and thus (8) diverges to $-\infty$ with probability one. Hence, under P_t^0, $dQ_t/dP_t^0 \to 0$ a.s. when $t \to \infty$. So we will choose the model without drift correctly with probability one as $t \to \infty$. Conversely, when $\theta \neq 0$, (8) is $(1/2t)X(t)^2 - (1/2)\ln(t)$ and this diverges to $+\infty$ a.s. when $t \to \infty$. Thus, under P_t^θ for $\theta \neq 0$, $dQ_t/dP_t^0 \to \infty$ a.s. when $t \to \infty$ and the model with drift is correctly chosen with probability one.

Section 4.4 below studies the case of vector θ from this point of view, considers the use of our formula as a model selection criterion in the general case, and explores its connection to BIC.

(iv) Theorem 2.1 remains true under random time changes. It is helpful to allow for such time changes because it is often convenient to measure time in terms of the amount of information that has accumulated about the process (i.e. A_t) rather than simply chronological time. Time changes are also useful in resolving integrability difficulties that sometimes arise in the formation of σ-finite measures like Q_t. For instance, the density process $\mathscr{R}_a$ given by the exponential in (7) above is not necessarily a proper probability density. This is because while $\mathscr{R}_a$ is a supermartingale it is not in general a martingale and it is not necessarily the case that $E(\mathscr{R}_a) = 1$. In fact, $\mathscr{R}_a$ is a martingale iff $E(\mathscr{R}_a) = 1$—see Karatzas and Shreve (1991, p. 198). The integrability problems arise when the quadratic variation $[G]_a$ that appears in the exponent of (7) becomes too large. This variation can be controlled by a suitable choice of stopping time sequence. In fact, it is known by Novikov's theorem (e.g. Karatzas and Shreve (1991, pp. 198–199)) that a sufficient condition for $E(\mathscr{R}_a) = 1$ is $E(\exp\{(1/2)[G]_a\}) < \infty$ and stopping times can be chosen to assure this. We now illustrate how to implement these ideas in our framework.

Suppose $(\tau_a)_{a>0}$ is a family of monotone increasing and continuous stopping times for which $\tau_a \to \infty$ as $a \to \infty$; then, in place of (3), we have for the time changed measures $\mathscr{P}_{\tau_a}$ and Q_{τ_a} a similar convergence as $a \to \infty$, i.e.

$$\frac{d\mathscr{P}_{\tau_a}}{dQ_{\tau_a}} = \frac{d\mathscr{P}_{\tau_a}}{dP^0_{\tau_a}} \Big/ \frac{dQ_{\tau_a}}{dP^0_{\tau_a}} \to 1 \text{ a.s. } (P^0), \quad \text{as } a \to \infty.$$

A convenient way of constructing stopping times that achieve the objective of making $\mathscr{R}_a$ a proper density is as follows. Set

$$\tau_a = \inf\{s : A_s \geq ce^a\}, \quad a \geq 0 \tag{9}$$

for some constant $c > 0$. The new initial time τ_0 can be interpreted as a minimum information time wherein we prescribe a level of minimum information, viz. c, that is needed for inference from the data to be useful. Note that at $t = 0$, $A_t = 0$ and (4) is undefined—we need some data (or $t > 0$ and $A_t > 0$) for the measure Q_t to be defined.

The sequence $(\tau_a)_{a \geq 0}$ is a family of monotone increasing and continuous (in a) stopping times such that A_{τ_a} is a.s. (P^0) bounded. The process is, in effect, stopped before the quadratic variation gets too large. We can go further and replace the time index t (chronological time) by a and let $a \to \infty$. This effects a time change in the process whereby the "new time" is measured by the information content of the original process. Correspondingly, we may replace the measure Q_t by the "time changed" exponential measure Q_a defined by

$$\frac{dQ_a}{dP} = \frac{dQ_{\tau_a}/dP^0_{\tau_a}}{dQ_{\tau_0}/dP^0_{\tau_0}} = \exp\{G_a + (1/2)[G]_a\}, \tag{10}$$

as in (7) above. In this new time frame and with the new initialization (at $a = 0$) Q_a as determined by (10) is a *proper* probability measure and it is *independent* of the prior distribution (or the level $\pi_0 = \pi(\theta^0)$ that appears in (4)). We formalize this result as follows.

2.4. THEOREM: *The measure Q_a that is defined by the RN derivative in* (10) *is a probability measure on the filtered space $(\Omega, \mathscr{F}_{\tau_a})$ for $a \geq 0$. This measure does not depend on the prior distribution $\pi(\theta)$, and*

$$Q_b|_{\mathscr{F}_{\tau_a}} = Q_a, \qquad \text{for all } \tau_b > \tau_a \geq \tau_0,$$

i.e. the restriction of Q_b to $\mathscr{F}_{\tau_a}$ is given by Q_a.

Thus, in the new time frame and with the new initialization the exponential Bayes measure Q_a is a proper probability measure and it is independent of the prior. In large samples we can therefore replace the Bayesian data measure by a probability measure of the general exponential form given in (10). In effect, for large samples we have left the prior density behind and have shown that only the score process V_t and its quadratic variation A_t are important in determining the

Bayesian data density. Our theorem shows that this holds under conditions which include nonstationary as well as stationary systems, and no rates of convergence conditions are required. Moreover, since Q_a is a proper probability measure it may be used to compute posterior odds and predictive odds and to compare models on the basis of these odds. We illustrate some of these uses in Section 4.

2.5. THE MODEL FOR Q_T: As shown in Remark 2.3(i) the measure Q_t gives rise to a proper probability density process $\mathscr{R}_a$ that can be represented in the exponential martingale form (7). This characterization of the density is useful in explaining how the Bayes factor dQ_t/dP_t^0 (or, more specifically, the conditional time changed Bayes factor $\mathscr{R}_a$) evolves as new data arrives.

Let $\hat{h}(t) = V(t)/A(t)$, which is, in effect, the linearized MLE (i.e., $\hat{\theta}_t - \theta^0 = V_t/A_t + o(\hat{\theta}_t - \theta^0)$ as $t \to \infty$—see equation (A17) in the Appendix). Then, $G_a = \int_{\tau_0}^{\tau_a} \hat{h}(t)\, dV(t)$. As in the proof of Theorem 2.4 in the Appendix, we find that $\mathscr{R}_a$ satisfies the equation

$$d\mathscr{R}_a = \mathscr{R}(\tau_a)\, dG(\tau_a) = \mathscr{R}(\tau_a)\hat{h}(\tau_a)\, dV(\tau_a). \tag{11}$$

This is a nonlinear stochastic differential equation for $\mathscr{R}_a = \mathscr{R}(\tau_a)$, showing how the density process $\mathscr{R}(\tau_a)$ is updated using the latest available (in $\mathscr{F}_{\tau_a}$) value of the linearized MLE $\hat{h}(\tau_a)$ and the increment in the score process $dV(\tau_a)$ at time τ_a. The model to which the exponential Bayes measure Q_t relates is therefore determined by the nonlinear stochastic differential equation (11), which prescribes the evolution of the path dependent density $\mathscr{R}_a = \mathscr{R}(\tau_a)$ from a given initialization at $a = 0$ (i.e. τ_0) in terms of the linearized MLE $\hat{h}(\tau_a)$ and the increment in the score $dV(\tau_a)$, both of which are continuously updated as a increases. Since Q_t is a large sample approximation to the Bayesian data measure $\mathscr{P}_t$ we can interpret (11) as a large sample approximation to the model *for the data* in a Bayesian framework. Note that (11) holds irrespective of the prior in large samples and so does this large sample approximating Bayes model.

With a further time change in the process it is possible to construct a more explicit model for the data that corresponds to the path dependent Bayes measure Q_t in the general case. To do this we use the following result:

2.6. LEMMA A: *Suppose V_t is a continuous local martingale with $V_0 = 0$ and quadratic variation process A_t for which $A_t \to \infty$ a.s. (P_0). Then there exists a Brownian motion X_t and a family of stopping times $(\sigma_t)_{t \geq 0}$ with $\sigma_t \to \infty$ as $t \to \infty$ such that V_t is indistinguishable from $\int_0^{\sigma_t} X\, dX$.*

The time change σ_t in the theorem is constructed using the rule

$$\sigma_t = \inf\left\{ p : \int_0^p X_s^2\, ds \geq A_t \right\}. \tag{12}$$

Then

$$\hat{h}_t = V_t/A_t = \int_0^{\sigma_t} X_s\, dX_s \Big/ \int_0^{\sigma_t} X_s^2\, ds, \tag{13}$$

$$G_a = G(\tau_a) = \int_{\tau_0}^{\tau_a} \hat{h}_d\, dV_t = \int_{\tau_0}^{\tau_a} \hat{h}_t X_{\sigma_t}\, dX_{\sigma_t}, \tag{14}$$

and

$$[G]_a = \int_{\tau_0}^{\tau_a} \hat{h}_t^2\, dA_t = \int_{\tau_0}^{\tau_a} \hat{h}_t^2 X_{\sigma_t}^2\, dt. \tag{15}$$

The time changed density process $\mathscr{R}_a$ now has the form of the exponential martingale

$$\mathscr{R}_a = \exp\{G_a - (1/2)[G]_a\} = \exp\left\{\int_{\tau_0}^{\tau_a} \hat{h}_t X_{\sigma_t}\, dX_{\sigma_t} - (1/2)\int_{\tau_0}^{\tau_a} \hat{h}_t^2 X_{\sigma_t}^2\, dt\right\}. \tag{16}$$

In fact, (16) is the likelihood ratio density process for the model (e.g., see Ibragimov and Has'minski (1981, p. 16)):

$$dX_{\sigma_t} = \hat{h}_t X_{\sigma_t}\, dt + dW_t, \quad t \geq \tau_0, \tag{17}$$

where W_t is a Brownian motion and $\hat{h}_t$ is given in (13). The nonlinear stochastic differential equation (17) is the model for the data that corresponds to the path dependent exponential Bayes measure Q_t. We can think of (17) as being the model for the score process V_t under the Bayes measure Q_t.

2.7. The Case of Quadratic Log Likelihood and Linear Diffusion: Let us now consider the special case of a quadratic log likelihood process. If also the prior density $\pi(\theta)$ is uniform, then the large sample approximations that appear in Theorem 2.1 are not needed and the analysis we have performed in the general case goes through exactly.

Suppose the log likelihood process is

$$l_t(\theta) = \ln(dP_t^\theta/dP_t^0) = V_t\theta - (1/2)A_t^2\theta, \tag{18}$$

corresponding to the linear diffusion model for the Ornstein-Uhlenbeck process

$$dX_t = \theta X_t\, dt + dW_t. \tag{19}$$

Then $V_t = \int_0^t X_s\, dX_s$ is a martingale under P^0 (i.e. when $\theta = 0$) and has quadratic variation $A_t = \int_0^t X_s^2\, ds$. The MLE of θ is given by $\hat{\theta}_t = V_t/A_t = \int_0^t X_s\, dX_t / \int_0^t X_s^2$ exactly and $\hat{\theta}_t = \hat{h}_t$ in the notation of Section 2.5.

The exponential Bayes measure Q_t now satisfies $d\mathscr{P}_t/dP_t^0 = dQ_t/dP_t$ exactly, and its conditional density process (17) is given by

$$q_a = dQ_{\tau_a}/dP_{\tau_a}^0|_{\mathscr{F}_{\tau_0}} = \exp\{G_a - (1/2)[G]_a\}, \tag{16'}$$

where

$$G_a = \int_{\tau_0}^{\tau_a} \hat{\theta}_t \, dV_t = \int_{\tau_0}^{\tau_a} \hat{\theta} \tau Y_t \, dY_t \tag{20}$$

and

$$[G]_a = \int_{\tau_0}^{\tau_a} \hat{\theta}_t^2 Y_t^2 \, dt. \tag{21}$$

As in the case of (16) and (17), we deduce from the form of (16′) that the model corresponding to Q_t is

$$dY_t = \hat{\theta}_t Y_t \, dt + dW_t, \quad t \geq \tau_0, \tag{19′}$$

conditional on Y_{τ_0} and where $W_t \equiv \mathrm{BM}(1)$. The model (19′) is the model for the data under the Bayes data measure Q_t and, like (17), this is a path dependent nonlinear stochastic differential equation. However, in the case of (19′) we do not need further time changes in the process (like those in Lemma A) to obtain this explicit representation, and the model holds exactly rather than as an asymptotic approximation.

3. THE DISCRETE TIME CASE AND AN EMBEDDING THEOREM

Let $Y^n = \{Y_t\}_1^n$ be a discrete time series defined on the filtered sequence of measurable spaces $\{\Omega, \mathcal{F}_n\}$. Let P_n^θ be a parameterized probability measure of Y^n with $\theta \in \mathbb{R}$. Suppose θ^0 is the true value of θ and that $P_n^\theta \ll \nu_n$, some σ-finite measure on $(\Omega, \mathcal{F}_n)$. We write the RN derivative of P_n^θ with respect to $P_n^0 = P_n^{\theta^0}$ as

$$L_n(\theta) = dP_n^\theta / dP_n^0 = (dP_n^\theta / d\nu_n) / (dP_n^0 / d\nu_n). \tag{22}$$

If $\pi(\theta)$ is a prior density on θ, then the Bayesian data measure is given by the mixture $\mathcal{P}_n = \int_{\mathbb{R}} \pi(\theta) P_n^\theta \, d\theta$, as in the continuous time case.

Let $l_n(\theta) = \ln(L_n(\theta))$, $l_n^{(1)}(\theta)$ be the score and $B_n(\theta) = \langle l_n^{(1)}(\theta) \rangle$ be the conditional quadratic variation. Set $L_0 = 1$, and write the log likelihood as the telescoping sum $l_n(\theta) = \ln(L_n(\theta)) = \sum_{k=1}^n \{\ln(L_k(\theta)) - \ln(L_{k-1}(\theta))\}$. Then the score function has the form:

$$l_n^{(1)}(\theta) = \sum_{k=1}^n (\partial / \partial\theta)[\ln(L_k(\theta)) - \ln(L_{k-1}(\theta))] = \sum_{k=1}^n \varepsilon_k(\theta), \quad \text{say,}$$

and

$$B_n(\theta) = \sum_{k=1}^n E\left(\varepsilon_k(\theta)^2 | \mathcal{F}_{k-1} \right) = \langle l_n^{(1)}(\theta) \rangle$$

is the conditional variance of the martingale $l_n^{(1)}(\theta)$ under P_n^θ (cf. Hall-Heyde (1980, p. 157)).

Under conditions (D1)–(D7) that are the mirror image in discrete time of (C1)–(C7) in continuous time (see the Appendix for an explicit statement of them), $\mathscr{P}_n$ can be asymptotically approximated as follows:

$$\frac{d\mathscr{P}_n}{dP_n^0} \Big/ \frac{dQ_n}{dP_n^0} \to 1 \text{ a.s. } (P^0). \tag{23}$$

Here Q_n is the measure defined by the following RN derivative with respect to P_n^0:

$$dQ_n/dP_n^0 = c_0 \exp\{(1/2)V_n^2 B_n^{-1}\}/B_n^{1/2}, \tag{24}$$

where $V_n = l_n^{(1)}(\theta^0)$, $B_n = B_n(\theta^0)$, and $c_0 = (2\pi)^{1/2}\pi_0$. The derivative (24) has the following asymptotically equivalent forms:

$$dQ_n/dP_n^0 = c_0 \exp\left\{(1/2)\left(\hat{\theta}_n - \theta^0\right)^2 B_n\right\} \Big/ B_n^{1/2} \tag{25}$$

and

$$dQ_n/dP_n^0 = c_0 \exp\left\{l_n\left(\hat{\theta}_n\right)\right\} \Big/ B_n^{1/2}. \tag{26}$$

3.1. EMBEDDING THE DISCRETE TIME DENSITY IN A CONTINUOUS PROCESS: It is rather more difficult than in the continuous time case to determine the form of the implied Bayes model from the form of the discrete time process (24). We can however use the theory for the continuous time case to analyze the discrete time case by an embedding technique. We will show that we can embed the process (24) into a corresponding continuous time process whose Bayes model we have already studied in Sections 2.5–2.7. The discrete time Bayes model can then be regarded as simply the model of the discrete observations from the continuous process. An advantage of this embedding is that we can analyze the model without making a special cut in the asymptotic theory for nonstationary time series (i.e. in the case of a unit root). This is because in the continuous time case there is no difference in treatment between the stationary and nonstationary cases.

To begin, we continue to assume conditions (D1)–(D7) hold and then the asymptotic approximation (23) applies. Our objective is to find an alternative representation of (24) in terms of a continuous process. It will be convenient for us to write the increments in the score process $l_n^{(1)}(\theta)$ at $\theta = \theta^0$ as $\varepsilon_k = \varepsilon_k(\theta^0)$. Then we have $V_n = l_n^{(1)} = \sum_{k=1}^n \varepsilon_k$, which is P_n^0-martingale with conditional variance process B_n. Let $\mathscr{F}_k$ be the σ-field generated by $(\varepsilon_j)_1^k$.

3.2. THEOREM: *Assume* (D1), (D2), *and the following conditions hold:*

(D8) $\sup_{k \geq 1} E(\varepsilon_k^4) < \infty$.

(D9) $\sup_{k \geq 1} E(\varepsilon_k^4 | \mathscr{F}_{k-1})/\{E(\varepsilon_k^2 | \mathscr{F}_{k-1})\}^2 \leq C_a$ *a.s.* (P^0) *for some constant* $C_a > 0$.

(D10) *There exists some* γ *with* $0 < \gamma < 1$ *such that*

$$\frac{E(\varepsilon_n^2|\mathscr{F}_{n-1})}{B_n^{\gamma}} \to 0 \ a.s.\ (P^0).$$

Then there exists a probability space $(\Omega, \mathscr{E}, P)$ *supporting* $(V_n, B_n)_{n\geq 1}$, *a standard Brownian motion* W, *and stopping times* $(\tau_n)_{n\geq 1}$ *such that*

$$\frac{\exp\{(1/2)V_n^2 B_n^{-1}\}}{B_n^{1/2}} \Bigg/ \frac{\exp\{(1/2)W(\tau_n)^2/\tau_n\}}{\tau_n^{1/2}} \to 1 \ a.s.\ (P^0). \tag{27}$$

3.3 REMARKS ON (D8)–(D10): (i) Condition (D8) requires that fourth moments of the martingale differences ε_k exist and are uniformly (in k) bounded above. It could be relaxed to a weaker $(2+r)$-moment requirement on ε_k for some r with $0 < r < 2$, at the expense of making the proof (and some of the other conditions) of Theorem 3.2 more complicated.

(ii) Condition (D9) imposes a bound on the relative conditional fourth moments of ε_k. (D9) requires that the ratio of the conditional fourth moment to the square of the conditional second moment of ε_k be uniformly bounded above. This means that the kurtosis of the conditional distribution of ε_k cannot be too large relative to the square of the variance. For a stochastic linear regression model $y_t = \theta' x_t + u_t$ with $u_t \equiv \text{iid } N(0,1)$ and $\mathscr{F}_{t-1}$-measurable regressors, the score process increments are $\varepsilon_k = x_k u_k$ and then

$$\sup_{k\geq 1} \frac{E(\varepsilon_k^4|\mathscr{F}_{k-1})}{\{E(\varepsilon_k^2|\mathscr{F}_{k-1})\}^2} = \sup_{k\geq 1} \frac{2\sigma^2 x_k^4}{\sigma^4 x_k^4} = 2.$$

In this case the condition (D9) is fulfilled regardless of the structure of the regressor x_t.

(iii) The conditional variance process B_n is often interpreted as the time clock of the martingale V_n in the sense that it records the information content of the process up to time period n. The increment in the information content from period $n-1$ to period n is

$$d_n = B_n - B_{n-1} = E(\varepsilon_n^2|\mathscr{F}_{n-1}).$$

Condition (D10) requires that the incremental information d_n be small (by an order of magnitude or power of B_n) relative to the total information content B_n. We can explore the implications of this requirement in the linear AR(1) model $y_t = \alpha y_{t-1} + u_t$, with $u_t \equiv \text{iid } N(0, \sigma^2)$. In this case we have $\varepsilon_k = y_{k-1} u_k$ and $E(\varepsilon_k^2|\mathscr{F}_{k-1}) = y_{k-1}^2 \sigma^2$. (D10) requires that

$$\frac{y_{n-1}^2 \sigma^2}{(\Sigma_1^n y_{k-1}^2 \sigma^2)^{\gamma}} \to 0 \text{ a.s. } (P), \tag{28}$$

for some γ in the interval $0 < \gamma < 1$. Take the stationary case first. Here $|\alpha| < 1$ and we have $n^{-1}\sum_1^n y_{k-1}^2 = 0_{\text{a.s.}}(1)$ and (28) holds if

$$y_{n-1}^2/n^\gamma \to 0 \text{ a.s. } (P),$$

which holds by the Borel Cantelli Lemma if $\sup_n E(y_n^4) < \infty$ and $\gamma > 1/2$. In the unit root case where $\alpha = 1$ we rescale the numerator and denominator of (28) as follows:

$$\frac{n \ln(\ln(n))\{y_{n-1}^2\sigma^2/n\ln(\ln(n))\}}{n^{2\gamma}/(\ln(\log(n)))^\gamma\{\sum_1^n y_{k-1}^2\sigma^2/[n^2/\ln(\ln(n))]\}^\gamma}. \tag{29}$$

By the law of the iterated logarithm we have

$$\limsup_{n\to\infty} \frac{y_{n-1}^2\sigma^2}{n\ln(\ln(n))} = 2\sigma^2 \text{ a.s. } (P),$$

and by a result of Donsker-Varadhan (1977, p. 751) that is used in Lai-Wei (1983, p. 364) we have

$$\liminf_{n\to\infty} \frac{\sum_1^n y_{n-1}^2\sigma^2}{n^2/\ln(\ln(n))} = \sigma^4/4 \text{ a.s. } (P),$$

so that (29) is of order $O(n^{-2\gamma+1}(\ln(\ln(n)))^{1+\gamma})$ and $\to 0$ a.s. (P) provided $\gamma > 1/2$. Hence, (28) and thus (D10) hold in the stationary and nonstationary AR(1) model for $\gamma > 1/2$.

3.4. REMARKS ON THEOREM 3.2: In the proof of Theorem 3.2 we use the fact that the discrete time martingale V_n can be embedded in a Brownian motion so that, by changing the probability space if necessary, we can write $V_n = W(\tau_n)$ a.s. (P) for some stopping time τ_n and a Brownian motion $W(t)$. This is simply an application of the conventional Skorokhod embedding of a martingale, as discussed in detail by Hall-Heyde (1980, Appendix 1). What Theorem 3.2 shows in addition is that it is possible at the same time to approximate the conditional variance process B_n by τ_n asymptotically. This means that the discrete data density

$$M_n = B_n^{-1/2}\exp\{(1/2)V_n^2/B_n\} \tag{30}$$

can be embedded asymptotically in the continuous process

$$R(t) = t^{-1/2}\exp\{(1/2)W(t)^2/t\} \tag{31}$$

using the stopping times τ_n. Following the analysis in Section 2, we now reinitialize the process $R(t)$ at some $t_0 > 0$ to avoid the discontinuity in the density at $t = 0$. The new initialization at t_0 also overcomes the more important problem of nonintegrability, discussed earlier in Remark 2.3(iv). For, the conditional distribution of $W(t)$ given $\mathcal{F}_{t_0}$ is $N(W(t_0, t-t_0))$ and therefore

$W(t)^2/(t-t_0)$, given $\mathcal{F}_{t_0}$, is noncentral chi-squared with one degree of freedom and noncentrality parameter $W(t_0)^2/(t-t_0)$. A simple calculation then reveals that

$$E\left(\exp\{(1/2t)W(t)^2\}|\mathcal{F}_{t_0}\right) = \exp\{(1/2t_0)W(t_0)^2\}(t/t_0)^{1/2} < \infty.$$

It follows that the reinitialized process $r(t) = R(t)/R(t_0)$ is integrable and, moreover, $E(r(t)|\mathcal{F}_{t_0}) = 1$, so that $r(t)$ is a proper probability density.

Proceeding as in Section 2 we can now write $r(t)$ in the exponential density form (cf. equation (7) above)

$$r(t) = \exp\{G(t) - (1/2)[G]_t\}, \tag{32}$$

where $G(t) = \int_{t_0}^{t}(W(s)/s)\,dW$ and $[G]_t = \int_{t_0}^{t}(W(s)^2/s^2)\,ds$. As in the derivation of (17), the model corresponding to $r(t)$ is then seen to be the nonlinear diffusion equation

$$dX(t) = \hat{h}_t\,dt + dW(t), \quad t > t_0 \tag{33}$$

where $\hat{h}_t = W(t)/t$ is the maximum likelihood estimator of the drift in the Wiener process $W(t)$, i.e. the parameter θ in the simple linear model

$$X(t) = \theta + W(t) \tag{34}$$

when the true $\theta = 0$.

Our embedding theory tells us that the Bayesian data density for the discrete time scalar parameter likelihood is (after reinitialization) asymptotically equivalent to appropriate discrete draws from the continuous process $r(t)$. But, as we have seen, the model for the data corresponding to $r(t)$ is the path dependent diffusion equation (33). In consequence, the model for the discrete data corresponding to the Bayesian measure Q_n defined by (24) is just appropriate discrete draws from the output of (33).

4. THE MULTIVARIATE CASE

In this section we consider the discrete time case as in Section 3 but allow for a vector of parameters $\theta \in \mathbb{R}^k$. In other respects the framework of Section 3 will stay the same. We will show that the Bayesian data measure $\mathcal{P}_n = \int_{\mathbb{R}^k}\pi(\theta)P_n^\theta\,d\theta$ can be approximated by an exponential measure, just as in the scalar parameter case. However, we want to proceed without having to be explicit about rates of convergence of individual components or linear combinations of the maximum likelihood estimator $\hat{\theta}_n$. This generality is helpful in models such as vector autoregressions with some cointegration and some unit roots because we do not then have to be specific about the directions in which cointegration occurs or the dimension of the cointegration space.

Our extension to the vector case involves some modifications to conditions (D2), (D3), (D4), and (D6) in the Appendix to accommodate the multivariate case. Our modified conditions are as follows:

(D2′) *Under* P_n^0, $l_n^{(1)}(\theta)$ *is a zero mean* L_2 *martingale with conditional quadratic variation (matrix) process* $B_n(\theta)$ *and* $\lambda_{\min}[B_n(\theta)] \to \infty$ *a.s.* (P^θ) *as* $n \to \infty$. *Let* $B_n = B_n(\theta^0)$.

(D3′) *Uniformly for* $h \in S_k = \{h \in \mathbb{R}^k : h'h = 1\}$

$$\{h' l_n^{(2)}(\theta) h + h' B_n(\theta) h\} / h' B_n(\theta) h \to 0 \text{ a.s. } (P^\theta) \text{ as } n \to \infty.$$

(D4′) *There exist continuous functions* $w_n(\theta, \theta')$ *such that* $w_n(\theta, \theta) = 0$ *and such that for some* $\delta > 0$ *and for all* $\theta, \theta' \in N_\delta(\theta^0) = \{\theta : \|\theta - \theta^0\| < \delta\}$ *we have*

$$\{h' l_n^{(2)}(\theta) h - h' l_n^{(2)}(\theta') h\} / h' B_n h \le w_n(\theta, \theta') \text{ a.s. } (P^0)$$

for each n *uniformly for* $h \in S_k$ *and* $w_n(\theta, \theta') \to w_\infty(\theta, \theta')$ *a.s.* (P^0) *uniformly for* $\theta, \theta' \in N_\delta(\theta)^0$.

(D6′) *For any* $\delta > 0$ *and* $\omega_\delta = \{\theta : \|\theta - \theta^0\| \ge \delta\}$ *we have*

$$|B_n|^{1/2} \int_{\omega_\delta} \pi(\theta)(dP_n^\theta / dP_n^0)\, d\theta \to 0 \text{ a.s. } (P^0).$$

4.1. THEOREM: *Under conditions* (*D*1), (*D*2′), (*D*3′), (*D*4′), (*D*5), (*D*6′), *and* (*D*7),

$$\frac{d\mathscr{P}_n}{dP_n^0} \Big/ \frac{dQ_n}{dP_n^0} \to 1 \text{ a.s. } (P^0) \tag{35}$$

where Q_n *is the exponential Bayes measure defined by the following RN derivative with respect to* P_n^0:

$$dQ_n / dP_n^0 = c_0 \exp\{(1/2) V_n' B_n^{-1} V_n\} \Big/ |B_n|^{1/2}, \tag{36}$$

where $V_n = l_n^{(1)}(\theta^0)$ *and* $c_0 = (2\pi)^{k/2} \pi_0$. *The following forms of the exponential density are asymptotically equivalent to* (36):

$$dQ_n / dP_n^0 = c_0 \exp\Big\{(1/2)\big(\hat\theta_n - \theta^0\big)' B_n \big(\hat\theta_n - \theta^0\big)\Big\} \Big/ |B_n|^{1/2}, \tag{37}$$

and

$$dQ_n / dP_n^0 = c_0 \exp\Big\{l_n\big(\hat\theta_n\big)\Big\} / |B_n|^{1/2}. \tag{38}$$

4.2. REMARKS ON THE NEW CONDITIONS: (i) (D2′) is just a vector version of (D2) and $\lambda_{\min}[B_n(\theta)] \to \infty$ a.s. (P^θ) corresponds to the usual excitation condition of regression theory. Similarly (D6′) is just the vector analogue of (D6).

(ii) (D3′) and (D4′) correspond to (D3) and (D4) but are written in terms of the quadratic forms $h' l_n^{(2)}(\theta) h$ and $h' B_n(\theta) h$ for a vector h on the unit sphere S_k in $\mathbb{R}^k$. In effect, $l_n^{(2)}(\theta) + B_n(\theta)$ must be uniformly small relative to $B_n(\theta)$ in all directions $h \in S_k$; and differences (measured in the direction h) between $l_n^{(2)}(\theta)$ and $l_n^{(2)}(\theta')$ relative to B_n must be bounded by the family $w_n(\theta, \theta')$ uniformly for $h \in S_k$.

4.3. REMARKS ON THEOREM 4.1: (i) When $\pi_0 = (2\pi)^{-k/2}$ we have $c_0 = 1$ and

$$(36') \qquad dQ_n/dP_n^0 = \exp\{(1/2)V_n' B_n^{-1} V_n\}/|B_n|^{1/2}.$$

This can be interpreted as a canonical form of the data density which depends only on the score process V_n and its conditional variance matrix B_n. Twice the logarithm of the likelihood ratio (36') is

$$2\ln(dQ_n/dP_n^0) = V_n' B_n^{-1} B_n - \ln|B_n|.$$

The first term in this expression is the score test of the hypothesis $\mathscr{H}_0 : \theta = \theta^0$. The second term is a penalty associated with the presence of the k free parameters in the vector θ and is discussed in the following section.

(ii) As in the univariate case, the posterior density is $\Pi_n^B(\theta) = \pi(\theta)\, dP_n^\theta / d\mathscr{P}_n$. This is asymptotically Gaussian of the form $N(\hat{\theta}_n, B_n^{-1})$, which is shown in the same way as Corollary 2.5.

(iii) Let $R_n = dQ_n/dP_n^0$ and suppose we condition on a minimal information time n_0. Then the large sample conditional data density at n_a given $\mathscr{F}_{n_0}$ is

$$(39) \qquad r_a = R_{n_a}/R_{n_0} = \frac{\exp\left\{(1/2)\left(V_{n_a}' B_{n_a}^{-1} V_{n_a} - V_{n_0}' B_{n_0}^{-1} V_{n_0}\right)\right\}}{|B_{n_a}|^{1/2}/|B_{n_0}|^{1/2}}$$

(we use the extra index "a" here to signify that a time change in the process may be performed to ensure integrability—cf. Remark 2.3(iv)). Note that the conditional density r_a given in (39) is independent of the prior density $\pi(\theta)$. We will not go into the details but it is possible to show, as we did in the continuous time case in Theorem 2.4, that r_a is a proper probability density.

(iv) Thus, the main result of Theorem 4.1 is that there is a generally applicable asymptotic theory which prescribes the form of the Bayesian data density as shown in (36) and this form allows for improper prior distributions. Moreover, although we may start with an improper prior distribution, our approximation (36) gives rise to proper probability densities as in (39) provided we condition on a minimal information time (like n_0) and stop the process (at n_a) if necessary to achieve integrability (as in the proof of Theorem 2.4 in the continuous time case). With these densities at hand in the convenient exponential form given, they can be used to compare models or to test hypotheses in terms of the relative impact of these hypotheses or model changes on the data density (i.e. by means of a likelihood ratio test or Bayes factor).

4.4. MODEL SELECTION AND THE RELATIONSHIP TO SCHWARZ'S (1978) BIC CRITERION: As indicated above, one consequence for practical work of our asymptotic theory is to the problem of model selection. We can use the exponential form of the data density to measure the evidence in the data in support of one model versus another. Our approach here is entirely analogous to that taken by Schwarz (1978) in "estimating the dimension of a model." Schwarz worked with iid observations from a distribution in the linear exponen-

tial family and adopted the Bayesian solution of selecting the model that is *a posteriori* the most probable in terms of its data density. In our approach we allow for a general log likelihood (like $l_n(\theta)$) and use the exponential data density asymptotic approximation as the basis of comparison between models, again selecting the one that is *a posteriori* most probable.

To fix ideas suppose we are given a general model for a time series $\{Y_n\}_1^n$ in terms of a parameterized probability measure P_n^θ with $\theta \in \Theta$ some convex set in $\mathbb{R}^k$. A class of competing models M_i $(i = 1, \ldots, I)$ is given in terms of the parameterized measures $P_n^{\theta_{k_i}}$ and the distinct parameter spaces $\Theta_{k_i} \ni \theta_{k_i}$, with $\Theta_{k_i} \subset \Theta$ and with $\dim(\Theta_{k_i}) = k_i$. Conditional on model i and given a prior $\pi_i(\theta_{k_i})$ for θ^{k_i} the Bayesian data measure is $\mathscr{P}_n^i = \int_{\Theta_{k_i}} \pi_i(\theta_{k_i}) P_n^{\theta_{k_i}} d\theta_{k_i}$. Let us now assume the existence of $\theta_{k_i}^0$ for which Theorem 4.1 holds for each $i = 1, \ldots, I$. Then the data density is approximated asymptotically by the exponential data density

$$q_{nk_i} = dQ_k^{k_i}/dP_n^0 = c_{0i} \exp\left\{ l_n\left(\hat{\theta}_{nk_i}\right)\right\} \Big/ |B_{ni}|^{1/2},$$

where $\hat{\theta}_{nk_i}$ is the maximum likelihood estimate of $\theta_{k_i}^0$, $l_n(\theta_{k_i}) = \ln\{L_n(\theta_{k_i})\}$ is the log-likelihood ratio function, $c_{0i} = (2\pi)^{k_i/2}\pi(\theta_{k_i}^0)$, and $B_{ni} = \langle l_n^{(1)}(\theta_{k_i}^0)\rangle$. Now

$$\ln(q_{nk_i}) = \ln\left(\hat{\theta}_{nk_i}\right) - \frac{1}{2}\ln|B_{ni}| + r_i,$$

where the remainder r_i is bounded as $n \to \infty$. Model selection proceeds by picking the model M_i which is the most likely given the observed data, i.e. the model that maximizes the (logarithm of the) exponential data density, eliminating the bounded remainder term r_i, viz.

$$PIC = \operatorname{argmax}_i\left[l_n\left(\hat{\theta}_{nk_i}\right) - (1/2)\ln|B_{ni}|\right]. \tag{40}$$

The criterion is denoted by "PIC" because it is a form of posterior information criterion. It can be compared with the so-called BIC criterion derived by Schwarz (1978, p. 461) for iid data in the linear exponential family, viz.

$$BIC = \operatorname{argmax}_i\left[l_n\left(\hat{\theta}_{nk_i}\right) - (k_i/2)\ln(n)\right]. \tag{41}$$

To relate these expressions, observe that $B_{ni} = \langle l_n^{(1)}(\theta_{k_i}^0)\rangle = \sum_{j=1}^n E(\varepsilon_j \varepsilon_j' | \mathscr{F}_{j-1})$ where $\varepsilon_j = \partial/\partial\theta[\ln L_n(\theta_{k_i})/L_{j-1}(\theta_{k_i}^0)]$, and for strictly stationary systems we can expect that

$$n^{-1}B_{ni} = n^{-1}\sum_{j=1}^n E\left(\varepsilon_j \varepsilon_j' | \mathscr{F}_{j-1}\right) \to_{\text{a.s.}} E(\varepsilon_j \varepsilon_j') = \Sigma_\varepsilon, \text{ say}$$

in which case we have

$$\begin{aligned}\ln|B_{ni}| &= \ln\{n^{k_i}|n^{-1}B_{ni}|\} = k_i \ln(n) + \ln|\Sigma_\varepsilon| + o_{as}(1) \\ &= k_i \ln(n) + o_{as}(1).\end{aligned}$$

Thus, in large samples and for stationary systems the criterion PIC is effectively equivalent to the criterion BIC.

In related work we have used the criterion PIC as an order selection procedure in autoregressive models with deterministic trends and compared its performance with the BIC criterion in finite samples in such models—see Phillips and Ploberger (1994) for details. The present analysis allows for more general models. It can, for example, be used to justify the joint order selection of lag length, trend degree, and cointegrating rank in a VAR model with deterministic trends and with potentially reduced rank (thereby allowing for cointegration). Some explicit results on this problem are contained in Phillips (1994) and Chao and Phillips (1994). In the much simpler univariate context we can use the method to assess sample evidence in support of the hypothesis of a unit root, as we now consider.

4.5. TESTING FOR THE PRESENCE OF A UNIT ROOT: We start with the simple model

$$Y_t = \alpha Y_{t-1} + u_t, \qquad u_t \equiv \text{iid } N(0, \sigma^2) \tag{42}$$

with σ^2 known, and the process in (1) initialized at $t = 0$ with Y_0 and $\mathcal{F}_0$-measurable variable. Since our interest is in the unit root hypothesis $\alpha = 1$ it is convenient to rewrite (42) as

$$\Delta Y_t = hY_{t-1} + u_t, \quad \text{with} \quad h = \alpha - 1. \tag{42'}$$

Let P_n^h be the probability measure of $Y^n = \{Y_t\}_1^n$ conditional on Y_0 and let $P_n = P_n^0$ be this measure when $h = 0$.

The exponential data density of Y^n is given by Theorem 3.1 in any of the equivalent forms (24)–(26). Using (24) we have

$$\begin{aligned} dQ_n/dP_n &= c_0 \exp\{(1/2)V_n^2 B_n^{-1}\}/B_n^{1/2} \\ &= c_0 \exp\left\{(1/2\sigma^2)\hat{h}_n^2 A_n\right\}/(A_n/\sigma^2)^{1/2} \end{aligned} \tag{43}$$

where $c_0 = (2\pi)^{1/2}\pi_0$ and

$$V_n = (1/\sigma^2)\sum_1^n Y_{t-1}\,\Delta Y_t,$$

$$B_n = \langle V_n \rangle = (1/\sigma^2)\sum_1^n Y_{t-1}^2 = (1/\sigma^2)A_n, \quad \text{say.}$$

We may treat the issue of whether or not to set $h = 0$ in (42′) as a model selection problem and use the PIC criterion given in (40). When $h = 0$ in (42′) the log likelihood ratio is $\ln(dP_n/dP_n) = 0$. When Y_{t-1} is included in the regression (42′) the log likelihood ratio is

$$\ln(dQ_n/dP_n) = (1/2\sigma^2)\hat{h}_n^2 A_n - (1/2)\ln(A_n/\sigma^2).$$

Thus, the criterion (40) becomes

$$PIC = \text{argmax}\left\{0, (1/2\sigma^2)\hat{h}_n^2 A_n - (1/2)\ln(A_n/\sigma^2)\right\} \tag{44}$$

and this is equivalent to the decision rule:

(R1) "*if*

$$\frac{dQ_n}{dP_n}(\sigma^2) = \frac{\exp\{(1/2\sigma^2)\hat{h}_n^2 A_n\}}{(A_n/\sigma^2)^{1/2}} > 1$$

then decide in favor of the model (42′) *over the model with a unit root* ($h = 0$)."

When σ^2 is unknown then we use the same rule (R1) but employ

$$\frac{dQ_n}{dP_n}(\hat{\sigma}^2) = \frac{\exp\{(1/2\hat{\sigma}^2)\hat{h}_n^2 A_n\}}{(A_n/\hat{\sigma}^2)^{1/2}}, \qquad \hat{\sigma}^2 = n^{-1}\sum_1^n \left(\Delta Y_t - \hat{h}_n Y_{t-1}\right)^2, \tag{45}$$

in place of $dQ_n/dP_n(\sigma^2)$. We call this procedure the PIC test for a unit root. Its asymptotic properties are analogous to those based on Bayes factors and are given in the following result.

4.6. THEOREM: *The PIC test for a unit root in the model* (42′) *is based on the decision rule*:

(R2) "*if* $dQ_n/dP_n(\hat{\sigma}^2) < 1$ *then accept the hypothesis of a unit root* (*i.e.* $h = 0$)."

This test is completely consistent in the sense that type I and type II errors both tend to zero as $n \to \infty$.

4.7. REMARKS ON THE PIC TEST: The decision rule (R2) is based on the model selection principle PIC. The PIC criterion (40) was obtained using the exponential data density (38). Now the precise form of $dQ_n/dP_n(\hat{\sigma}^2)$ that is given in (45) and used in (R2) is the same as the canonical form of the exponential data density—see Remark 4.3(i). This canonical form sets the multiplicative constant c_0 that appears in the density (38) (or (24) in the univariate case here) to the specific value $c_0 = 1$. If $\pi(h)$ is the prior density of the parameter h in (42′), then use of the canonical form with $c_0 = 1$ is equivalent to setting $\pi_0 = (2\pi)^{-1/2}$ as the value of the prior at $h = 0$. Note that this setting of π_0 does not mean that the prior itself has to be uniform and set at this level (although this certainly could be the case). In view of the asymptotic theory given in Theorem 4.1, the requirement behind (24) is only that the prior $\pi(h)$ be continuous at $h = 0$. For $c_0 = 1$ in (24), we then also need $\pi(0) = (2\pi)^{-1/2}$, as would be the case, for example, if the prior were $\pi(h) \equiv N(0, 1)$, which we might think of as a canonical prior for h in (42′). Clearly, results of tests that are based on the decision rule (R2) may be sensitive to the particular setting $c_0 = 1$ (or $\pi_0 = (2\pi)^{-1/2}$) that we have used in the construction of the PIC statistic $dQ_n/dP_n(\hat{\sigma}^2)$. This dependence on $\pi(h)$ is inevitable if we wish to use all of the data in the sample. But there is an alternative if we want to be independent of the prior and are prepared to give up some data points. We now consider this alternative.

For some $n_0 \geq 1$ we may form the conditional density of the measure $Q_n(\cdot|\mathcal{F}_{n_0})$ from the ratio

$$(46) \qquad r_n(\sigma^2) = \frac{dQ_n/dP_n}{dQ_{n_0}/dP_{n_0}}(\sigma^2) = \frac{\exp\left\{(1/2\sigma^2)\left[\hat{h}_n^2 A_n - \hat{h}_{n_0}^2 A_{n_0}\right]\right\}}{(A_n/A_{n_0})^{1/2}}$$

just as in (39) above. We may interpret n_0 as a minimal information time. For instance, when $n_0 = 1$ there is just enough data to estimate h in (42′) by $\hat{h}_1 = Y_0\,\Delta Y_1/Y_0^2 = \Delta Y_1/Y_0$. We can then use the common data set over $n_0 + 1 \leq t \leq n$ to compare two models (i.e. with and without the unit root) using the density $r_n(\sigma^2)$, or if σ^2 is unknown, $r_n(\hat{\sigma}^2)$. Note that $r_n(\sigma^2)$ and $r_n(\hat{\sigma}^2)$ do not depend on c_0 and are, in fact, independent of the prior distribution $\pi(h)$. Thus, by conditioning on the initial data over $0 \leq t \leq n_0$ we end up with a conditional form of the exponential data density that is independent of the prior and can be used for statistical testing. In place of (R2) we have the following decision rule:

(R3) "*if* $r_n(\hat{\sigma}^2) < 1$ *then accept the hypothesis of a unit root* (*i.e.* $h = 0$)."

We call the test based on $r_n(\hat{\sigma}^2)$ a conditional PIC test. If we are concerned about sensitivity of the PIC test outcome to the canonical factor $c_0 = 1$, then we can use the conditional PIC test and rule (R3) instead of (R2).

4.8 A USEFUL ALGEBRAIC FORM OF $r_n(\sigma^2)$: Using recursive least squares algebra (see Brown, Durbin, and Evans (1975)) we obtain a useful alternative form of $r_n(\sigma^2)$.

LEMMA B:

$$(47) \qquad r_n(\sigma^2) = \prod_{t=n_0+1}^{n} \frac{(1/2\pi f_t)^{1/2}\exp\left\{-(1/2f_t)\left(\Delta Y_t - \hat{h}_{t-1}Y_{t-1}\right)^2\right\}}{(1/2\pi)^{1/2}\exp\left\{-(1/2)(\Delta Y_t)^2\right\}}$$

$$\text{with } f_t = \sigma^2(1 + Y_{t-1}^2/A_{n-1})$$

and

$$(48) \qquad dQ_n/dQ_{n_0} = \prod_{t=n_0+1}^{n} (1/2\pi f_t)^{1/2}\exp\left\{-(1/2f_t)\left(\Delta Y_t - \hat{h}_{t-1}Y_{t-1}\right)^2\right\}.$$

Expression (47) is useful because it shows exactly how the density $r_n(\sigma^2)$ is constructed on a period by period basis. Using the fact that $dP_n/dP_{n_0} = \prod_{n_0+1}^{n}[(1/2\pi\sigma^2)^{1/2}\exp\{-(1/2\sigma^2)(\Delta Y_t)^2\}]$, expression (47) and the definition of $r_n(\sigma^2)$ give (48), which is the conditional density of the measure Q_n given $\mathcal{F}_{n_0}$. Since $dQ_n/dQ_{n_0} = (dQ_n/d\nu)/(dQ_{n_0}/d\nu)$, we see that (48) is, in fact, the conditional density with respect to Lebesgue measure (ν) of Q_n given $\mathcal{F}_{n_0}$. The form

of (48) also reveals that it is the density of data from the model

$$\Delta Y_t = \hat{h}_{t-1} Y_{t-1} + v_t, \quad \text{for} \quad t \geq n_0 + 1 \tag{49}$$

with $v_t \equiv$ i.ni.d $N(0, f_t)$. Thus, in this simple Gaussian case, the path dependent model that corresponds to the data measure Q_n can be obtained in the explicit form of (49). This result is much simpler than the general case studied in Section 3, where we need to embed the density of Q_n in a continuous time process in order to analyze the path dependent model for the data. Note that (49) is a predictive model for the data. The conditional PIC criterion $r_n(\sigma^2)$ given in (47) is, in fact, a predictive odds criterion for comparing the model (49) with the unit root model that has no estimated regression coefficient. This interpretation is explored further in Phillips and Ploberger (1994) and Phillips (1994).

5. CONCLUSION

This paper is a beginning. It provides the limiting form of the Bayesian data density for a general case of likelihoods and prior distributions. The limit formula is an exponential density that depends on the score process and its conditional variance matrix. In large samples and when we condition on the data, the prior distribution is effectively washed out, so that the score process and its conditional variance matrix are the only factors that determine the behavior of the data density. These factors are the common elements in a fairly wide class of problems. To this extent, we can say that in large samples a single theory based on the exponential data density is possible in a Bayesian analysis. The practical applications of this theory that we have given here are to problems of model selection and unit root testing. Some further applications of the theory are given in Phillips and Ploberger (1994) to ARMA models with trends and unit roots and in Phillips (1994) to reduced rank vector autoregressions and Bayesian vector autoregressions.

Cowles Foundation for Research in Economics, Dept. of Economics, P.O. Box 208281, New Haven, CT 06520, U.S.A.

and

Department of Economics, University of St. Andrews, Fife, KY169AC, Scotland, U.K.

Manuscript received March, 1991; final revision received February, 1995.

APPENDIX

PROOF OF THEOREM 2.1: The proof follows the general idea given in Walker (1969) and Hartigan (1983, Sec. 11.2), but does not rely on a specific rate of convergence for the MLE $\hat{\theta}_t$, nor on asymptotic normality of $\hat{\theta}_t$, nor on any ergodic properties for the Fisher information.

As in (C6) define $\omega_\delta = \{\theta : |\theta - \theta^0| \geq \delta > 0\}$ and let $N_\delta = \mathbb{R} - \omega_\delta$. We can choose $\delta > 0$ such that N_δ corresponds to the neighborhood of θ^0 in (C4), i.e. $N_\delta(\theta^0)$. We can also choose $\delta = \delta(\varepsilon)$, given

some $\varepsilon > 0$, in such a way that

$$\text{(A1)} \qquad 1 - \varepsilon < \inf_{\theta \in N_\delta} \frac{\pi(\theta)}{\pi(\theta_0)} \leq \sup_{\theta \in N_\delta} \frac{\pi(\theta)}{\pi(\theta_0)} < 1 + \varepsilon$$

in view of condition (C7).

We have

$$\text{(A2)} \qquad d\mathscr{P}_t / dP_t^0 \left(\int_{N_\delta} + \int_{\omega_\delta} \right) \pi(\theta)(dP_t^\theta / dP_t^0)\, d\theta = I_\delta + I_\delta^c, \quad \text{say}$$

and by (C6)

$$\text{(A3)} \qquad A_t^{1/2} I_\delta^c \to 0 \text{ a.s. } (P^0).$$

Next write I_δ as

$$I_\delta = \int_{N_\delta} \pi(\theta)(dP_t^\theta / dP_t^0)\, d\theta = \int_{N_\delta} \pi(\theta) \exp\{l_t(\theta)\}\, d\theta,$$

and define for some large $M > 0$ the shrinking neighborhood of $\hat{\theta}_t$

$$N_t = \left\{ \theta : (\theta - \hat{\theta}_t)^2 A_t < M \right\},$$

with $N_t^c = \mathbb{R} - N_t$. Then

$$\text{(A4)} \qquad I_\delta = \int_{N_\delta \cap N_t} + \int_{N_\delta \cap N_t^c} J\pi(\theta) \exp\{l_t(\theta)\}\, d\theta = [I_1 + I_2], \quad \text{say.}$$

Consider I_1 first. Taking a second order Taylor expansion of $l_t(\theta)$ we have

$$\text{(A5)} \qquad l_t(\theta) = l_t(\hat{\theta}_t) + (1/2) l_t^{(2)}(\theta_m)(\theta - \hat{\theta}_t)^2$$

where θ_m lies on the line segment between $\hat{\theta}_t$ and θ. Now

$$\text{(A6)} \qquad l_t^{(2)}(\theta_m)(\theta + \hat{\theta}_t)^2 = -A_t(\theta - \hat{\theta}_t)^2 + \{[l_t^{(2)}(\theta_m) - l_t^{(2)}(\theta^0)]/A_t + [l_t^{(2)}(\theta^0) + A_t]/A_t\}(\theta - \hat{\theta}_t)^2 A_t.$$

Under (C3)

$$\text{(A7)} \qquad [l_t^{(2)}(\theta_0) + A_t]/A_t \to 0 \text{ a.s. } (P^0),$$

and under (C4)

$$\text{(A8)} \qquad |l_t^{(2)}(\theta_m) - l_t^{(2)}(\theta^0)|/A_t \leq w_t(\theta_m, \theta^0) \to 0 \text{ a.s. } (P^0),$$

uniformly for $\theta \in N_\delta$. Hence combining (A5)–(A8) we have

$$l_t(\theta) = l_t(\hat{\theta}_t) - (1/2) A_t(\theta - \hat{\theta}_t)^2 [1 + \varepsilon_{1t}(\theta)]$$

where $\varepsilon_{1t}(\theta) \to 0$ a.s. (P^0) uniformly for $\theta \in N_\delta$. Also $\pi(\theta) = \pi(\theta^0) + o_{as}(1)$ uniformly for $\theta \in N_\delta \cap N_t$ in view of (A1) and the definition of N_t. Using these expansions we have

$$\text{(A9)} \qquad I_1 = \exp\{l_t(\hat{\theta}_t)\} \int_{N_\delta \cap N_t} [\pi_0 + \varepsilon_{1t}(\theta)] \exp\left\{ -(1/2) A_t(\theta - \hat{\theta}_t)^2 [1 + \varepsilon_{1t}(\theta)] \right\} d\theta$$

$$= A_t^{-1/2} \exp\{l_t(\hat{\theta}_t)\} (2\pi)^{1/2} \pi_0 [1 + O(\exp(-M/2)) + O(\eta_t)],$$

where for $\theta \in N_\delta$ we have $|\varepsilon_{it}(\theta)| \leq \eta_t \to 0$ a.s. (P^0) for $i = 1, 2$. It is in fact possible to choose M in the definition of N_t in such a way that $M \to \infty$ as $t \to \infty$. We may, for instance, choose $M = M_t =$

$(1/2)\delta^2 A_t$, and then

(A10) $\qquad M_t \to \infty$ a.s. (P^0) as $t \to \infty$.

Now consider I_2 in (A4). Using (A5) again we have

$$I_2 = \int_{N_\delta \cap N_t^c} \pi(\theta)\exp\{l_t(\theta)\}\,d\theta = \exp\{l_t(\hat\theta_t)\}\int_{N_\delta \cap N_t^c} \pi(\theta)\exp\left\{(1/2)l_t^{(2)}(\theta_m)(\theta-\hat\theta_t)^2\right\}d\theta.$$

Now

$$l_t^{(2)}(\theta_m) = -A_t\{1 - [A_t + l_t(\theta^0)]/A_t - [l_t^{(2)}(\theta_m) - l_t(\theta^0)]/A_t\}$$

and in view of (A7) and (A8) we find that for large enough t

$$l_t^{(2)}(\theta_m) < -(1/2)A_t \text{ a.s. } (P^0)$$

for $\theta \in N_\delta$. It follows that we may bound I_2 by the expression

(A11)
$$\begin{aligned} I_2 &\le (1+\varepsilon)\pi_0\exp\{l_t(\hat\theta_t)\}\int_{N_\delta \cap N_t^c}\exp\left\{-(1/4)A_t(\theta-\hat\theta_t)^2\right\}d\theta \\ &\le (1+\varepsilon)\pi_0\exp\{l_t(\hat\theta_t)\}\int_{N_t^c}\exp\left\{-(1/4)A_t(\theta-\hat\theta_t)^2\right\}d\theta \\ &= (1+\varepsilon)\pi_0 A_t^{-1/2}\exp\{l_t(\hat\theta_t)\}(2\pi)^{1/2}O(\exp\{-(1/4)M\}). \end{aligned}$$

Combining (A9) and (A11) we have

(A12) $\qquad I_\delta = (2\pi)^{1/2}\pi_0 A_t^{-1/2}\exp\{l_t(\hat\theta_t)\}[1 + o_{as}(1)]$

and then, using (A1), (A2), and (A12) we obtain

(A13) $\qquad d\mathscr{P}_t/dP_t^0 = I_\delta + I_\delta^c = (2\pi)^{1/2}\pi_0 A_t^{-1/2}\exp\{l_t(\hat\theta_t)\}[1 + o_{as}(1)].$

To complete the proof of the theorem we find an alternative representation of the factor $\exp\{l_t(\hat\theta_t)\}$ in (A13). Noting that $l_t(\theta^0) = 0$ we have the two Taylor expansions

(A14) $\qquad l_t(\hat\theta_t) = l_t^{(1)}(\theta^0)(\hat\theta_t - \theta^0) + (1/2)l_t^{(2)}(\theta_{m_1})(\hat\theta_t - \theta^0)^2$

and

(A15) $\qquad 0 = l_t^{(1)}(\hat\theta_t) = l_t^{(1)}(\theta^0) + l_t^{(2)}(\theta_{m_2})(\hat\theta_t - \theta^0)$

with θ_{m_1} and θ_{m_2} lying on the line segment joining $\hat\theta_t$ and θ^0. Combining (A14) and (A15) we have

$$\begin{aligned} l_t(\hat\theta_t) &= (1/2)(\hat\theta_t - \theta^0)^2\{l_t^{(2)}(\theta_{m_1}) - 2l_t^{(2)}(\theta_{m_2})\} \\ &= (1/2)(\hat\theta_t - \theta^0)^2 A_t\{[l_t^{(2)}(\theta_{m_1}) - l_t^{(2)}(\theta^0) + l_t^{(2)}(\theta^0) + A_t]/A_t \\ &\quad - 2[l_t^{(2)}(\theta_{m_2}) - l_t^{(2)}(\theta^0) + l_t^{(2)}(\theta^0) + A_t]/A_t + 1\} \\ &= (1/2)(\hat\theta_t - \theta^0)^2 A_t[1 + o_{as}(1)], \end{aligned}$$

using (C3) and (C4). It follows that (A13) may also be written as

(A16) $\qquad d\mathscr{P}_t/dP_t^0 = (2\pi)^{1/2}\pi_0 A_t^{-1/2}\exp\left\{(1/2)(\hat\theta_t - \theta^0)^2 A_t\right\}[1 + o_{as}(1)],$

giving the stated result (5).

Finally, we can use (A15) again, giving

$$\text{(A17)} \qquad 0 = l_t^{(1)}(\theta^0) + \{l_t^{(2)}(\theta_{m_2}) - l_t^{(2)}(\theta^0) + l_t^{(2)}(\theta^0) + A_t - A_t\}(\hat{\theta}_t - \theta^0)$$

$$= l_t^{(1)}(\theta^0) - A_t(\hat{\theta}_t - \theta^0)[1 + o_{as}(1)]$$

in view of (C3) and (C4). Noting that $l_t^{(1)}(\theta^0) = V_t$ is a P_t^0 martingale, we can combine (A17) and (A16) to give

$$\text{(A18)} \qquad d\mathscr{P}_t/dP_t^0 = (2\pi)^{1/2}\pi_0 A_t^{-1/2}\exp\{(1/2)V_t^2 A_t^{-1}\}[1 + o_{as}(1)]$$

as required by expression (4). Using all three asymptotically equivalent forms of dQ_t/dP_t^0 given by (4), (5), and (6) we have

$$\frac{d\mathscr{P}_t}{dP_t^0} \Big/ \frac{dQ_t}{dP_t^0} \to 1 \text{ a.s. } (P^0)$$

and the theorem is proved. *Q.E.D.*

PROOF OF THEOREM 2.4: We start by writing

$$\text{(A19)} \qquad \mathscr{R}_a = \exp\{K(\tau_a) - K(\tau_0)\} = \exp\left\{\int_{\tau_0}^{\tau_a} dK(t)\right\},$$

which we note is independent of the prior distribution $\pi(\theta)$. The stochastic differential $dK(t)$ that appears in the last expression of (A19) can be evaluated by applying Ito calculus to $K(t) = V(t)^2/A(t) - (1/2)\ln(A(t))$. We obtain

$$\text{(A20)} \qquad dK(t) = [V(t)/A(t)]\,dV(t) - (1/2)[V(t)/A(t)]^2\,dA(t)$$

and using (A20) in (A19) we deduce that

$$\text{(A21)} \qquad \mathscr{R}_a = \exp\{G_a - (1/2)[G]_a\}$$

where $G_a = G(\tau_a) = \int_{\tau_0}^{\tau_a}[V(t)/A(t)]\,dV(t)$.

Since $V(t)$ is a martingale, $G(t) = \int_{\tau_0}^{t}[V(s)/A(s)]\,dV(s)$ is a martingale also and its quadratic variation process is $[G](t) = \int_{\tau_0}^{t}[V(s)/A(t)]^2\,dA(s)$. This gives us the exponential process (the so-called Doléans exponential—see Meyer (1989, p. 148)):

$$\mathscr{R}(t) = \exp\{G(t) - (1/2)[G](t)\}.$$

The process $\mathscr{R}_a$ in (A21) is obtained from $\mathscr{R}(t)$ by using the stopping times τ_a, i.e. $\mathscr{R}_a = \mathscr{R}(\tau_a)$. In view of the construction of the sequence τ_a (see (9)) $A(t)$, and hence $G(t)$, are bounded a.s. (P^0) in $\tau_0 \le t \le \tau_a$, so that $E[\exp\{(1/2)[G]_a\}] < \infty$. It follows by Novikov's Theorem (e.g., see Ikeda and Watanabe (1989, Theorem 5.3, p. 152)) that $E[\mathscr{R}_a] = 1$ and $\mathscr{R}_a$ is therefore a continuous L_2 martingale.

It follows that the measure Q_a that is determined by the *RN* derivative $dQ_a/dP = \mathscr{R}_a$ is a proper probability measure, with probabilities given by integrals of $\mathscr{R}_a$, viz.

$$Q_a(B) = \int_B \mathscr{R}_a\,dP, \qquad \forall B \in \mathscr{F}_{\tau_a},$$

and $Q_b|\mathscr{F}_{\tau_a} = Q_a$ for all $\tau_b > \tau_a \ge \tau_0$, as in Ikeda and Watanabe (1989, p. 191). *Q.E.D.*

ASYMPTOTIC NORMALITY OF THE POSTERIOR: Work in the framework of Section 2 and define the posterior density process for θ by the ratio (using Bayes rule)

$$\begin{aligned}\Pi_t^B(\theta) &= \pi(\theta)(dP_t^\theta/dP_t^0)\Big/\int_{\mathbb{R}}\pi(\theta)(dP_t^\theta/dP_t^0)\,d\theta \\ &= \pi(\theta)(dP_t^\theta/dP_t^0)/(d\mathscr{P}_t/dP_t^0) \\ &= \pi(\theta)\,dP_t^\theta/d\mathscr{P}_t.\end{aligned}$$

Applying Theorem 2.1 we see that the Bayes data measure $\mathscr{P}_t$ in this expression for $\Pi_t^B(\theta)$ can be replaced by the exponential Bayes measure Q_t with a relative error that tends to zero as $t \to \infty$, i.e. the asymptotic form of the posterior density process is simply

$$\Pi_t^B(\theta) \sim \pi(\theta)(dP_t^\theta/dQ_t), \quad \text{as} \quad t \to \infty.$$

As the following Corollary to Theorem 2.1 shows, the density $\Pi_t^B(\theta)$ is, in fact, asymptotically Gaussian in form with a $N(\hat{\theta}_t, A_t^{-1})$ density. The asymptotic form of $\Pi_t^B(\theta)$ given above shows that the asymptotic Gaussianity of $\Pi_t^B(\theta)$ should be interpreted in the light of the reference measure Q_t with respect to which the likelihood (viz. dP_t^θ/dQ_t) is implicitly being computed. In effect, this change of reference measure from P_t^0 to Q_t alters the frame of reference (of model) with respect to which that Gaussian posterior density for θ should be interpreted.

COROLLARY TO THEOREM 2.1: *Suppose the conditions of Theorem* 2.1 *hold. Given* $M > 0$, *let* $N_t^M = \{\theta : (\theta - \hat{\theta}_t)^2 A_t \le M\}$ *and define*

$$\varphi(\theta; \hat{\theta}_t, A_t^{-1}) = (2\pi)^{-1/2} A_t^{1/2} \exp\left\{-(1/2)(\theta - \hat{\theta}_t)^2 A_t\right\}.$$

The posterior density $\Pi_t^B = \pi(\theta)\,dP_t^\theta/d\mathscr{P}_t$, *is asymptotically Gaussian* $N(\hat{\theta}_t, \hat{A}_t^{-1})$ *with density* (13) *in the sense that*

$$\sup_{\theta \in N_t^M}\left|\frac{\Pi_t^B(\theta)}{\varphi(\theta; \hat{\theta}_t, A_t^{-1})} - 1\right| \to 0 \text{ a.s. } (P^0)$$

as $t \to \infty$.

PROOF OF COROLLARY TO THEOREM 2.1: Using the same line of argument as that leading up to (A9) in the proof of Theorem 2.1 we have

$$\begin{aligned}\Pi_t^B(\theta) &= \pi(\theta)(dP_t^\theta/d\mathscr{P}_t) = \pi(\theta)(dP_t^\theta/dP_t^0)/(d\mathscr{P}_t/dP_t^0) \\ &= \pi(\theta)(dP_t^\theta/dP_t^0)/(dQ_t/dP_t^0)[1 + o_{as}(1)] \\ &= (2\pi)^{-1/2}(\pi(\theta)/\pi_0)A_t^{1/2}\exp\{l_t(\theta) - l_t(\theta) - l_t(\hat{\theta}_t)\}[1 + o_{as}(1)] \\ &= (2\pi)^{-1/2}[1 + \varepsilon_{2t}(\theta)]A_t^{1/2}\exp\left\{-(1/2)(\theta - \hat{\theta}_t)^2 A_t[1 + \varepsilon_{1t}(\theta)][1 + o_{as}(1)]\right\}\end{aligned}$$

where $\varepsilon_{it}(\theta) \to$ a.s. (P^0) uniformly in $N_\delta \cap N_t^M$ for $i = 1, 2$. Since $\hat{\theta}_t \to_{\text{a.s.}} \theta^0$ and $A_t = A_t(\theta^0) \to \infty$ a.s. (P^0) we have $N_t^M \subset N_\delta$ a.s. (P^0) for large enough t and fixed $M > 0$. Then

$$\sup_{\theta \in N_t^M}\left|\frac{\Pi_t^b(\theta)}{\varphi(\theta; \hat{\theta}_t, A_t^{-1})} - 1\right| \to 0 \text{ a.s. } (P^0),$$

giving (14), as required.

PROOF OF LEMMA 1: Under the state conditions it is well known that there is a stopping time

$$\nu_t = \inf\{s : A_s \ge t\}$$

such that V_{ν_t} is indistinguishable from a Brownian motion W_t (e.g., Protter (1990), Theorem 41, p.

81). We can write this equivalence as $V_t = W_{A_t}$ a.s. $0 \le t < \infty$. Now let X_t be another Brownian motion on the space and construct a new family of stopping times $(\sigma_t)_{t \ge 0}$ as

$$\sigma_t = \inf\left\{ p : \int_0^p X_s^2 \, ds \ge A_t \right\}.$$

Then $\int_0^{\sigma_t} X dX$ is a martingale with quadratic variation process $\int_0^{\sigma_t} X_s^2 \, ds = A_t$ a.s. $0 \le t < \infty$. Like V_t, the process $\int_0^{\sigma_t} X dX$ is equivalent to the time changed Brownian motion W_{A_t}. Hence, we have

$$V_{\nu_t} = \int_0^{\sigma_{\nu_t}} X dX = W_t \text{ a.s.} \quad 0 \le t < \infty,$$

and

$$V_t = \int_0^{\sigma_t} X dX = W_{A_t} \text{ a.s.} \quad 0 \le t < \infty,$$

giving the required result.

REGULARITY CONDITIONS FOR THE DISCRETE TIME CASE:

(D1) *$l_n(\theta) = \ln(L_n(\theta))$ is twice continuously differentiable with derivatives $l_n^{(1)}(\theta)$ and $l_n^{(2)}(\theta)$.*

(D2) *Under P_n^θ, $l_n^{(1)}(\theta)$ is a zero mean L_2 martingale with conditional quadratic variation process $B_n(\theta)$ and $B_n(\theta) \to \infty$ a.s. (P^θ) as $n \to \infty$. Let $B_n = B_n(\theta^0)$.*

(D3) *$(l_n^{(2)}(\theta))/B_n(\theta) \to 0$ a.s. (P^θ) as $n \to \infty$.*

(D4) *There exist continuous functions $w_n(\theta, \theta')$ such that $w_n(\theta, \theta) = 0$ and such that for some $\delta > 0$ and for all $\theta, \theta' \in N_\delta(\theta^0) = \{\theta : |\theta - \theta^0| < \delta\}$ we have*

$$\{l_n^{(2)}(\theta) - l_n^{(2)}(\theta')\}/B_n \le w_n(\theta, \theta') \text{ a.s. } (P^0)$$

for each n and $w_n(\theta, \theta') \to w_\infty(\theta, \theta')$ a.s. (P^0) uniformly for $\theta, \theta' \in N_\delta(\theta^0)$ and $w_\infty(\theta, \theta) = 0$.

(D5) *The maximum likelihood estimate $\hat{\theta}_n \to \theta^0$ a.s. (P^0).*

(D6) *For any $\delta > 0$ and $\omega_\delta = \{\theta : |\theta - \theta^0| \ge \delta\}$ we have*

$$B_n^{1/2} \int_{\omega_\delta} \pi(\theta)(dP_n^\theta / dP_n^0) \, d\theta \to 0 \text{ a.s. } (P^0).$$

(D7) *The prior density $\pi(\theta)$ is continuous at θ^0 with $\pi_0 = \pi(\theta^0) > 0$.*

PROOF OF THE ASYMPTOTIC APPROXIMATION (23): The proof is virtually identical to the proof of Theorem 2.1 but uses the conditional variance process $B_n = \langle l_n^{(1)}(\theta^0) \rangle$ in place of the quadratic variation $A_t[l_t^{(1)}(\theta_0)]$.

PROOF OF THEOREM 3.4: Since $\{V_n, \mathscr{F}_n, n \ge 1\}$ is a zero mean L_2 martingale we can embed this process in a standard Brownian motion. By Theorem A1, p. 269 of Hall-Heyde (1980) there exists probability space $(\Omega, \mathscr{E}, P)$ supporting $(V_n = \Sigma_1^n \varepsilon_k)_{n \ge 1}$, a standard Brownian motion W and stopping times $(\tau_n)_{n \ge 1}$ such that $V_n = W(\tau_n)$ and, if $\mathscr{E}_n \subset \mathscr{E}$ is the σ-field generated by $(V_k)_1^n$ and $W(t)$ for $0 \le t \le \tau_n$, then:

H^2(i) τ_n is $\mathscr{E}_n$-measurable:

H^2(ii) $E\{(\tau_n - \tau_{n-1})^2 | \mathscr{E}_{n-1}\} \le C_2 E(\varepsilon_n^4 | \mathscr{E}_{n-1})$ a.s. (P) where $C_2 = 16/\pi^2$; and

H^2(iii) $E\{(\tau_n - \tau_{n-1}) | \mathscr{E}_{n-1}\} = E(\varepsilon_n^2 | \mathscr{E}_{n-1})$ a.s. (P).

To prove (27) we need to show that

$$\text{(A22)} \qquad [V_n^2/B_n - W(\tau_n)^2/\tau_n] - \ln(B_n/\tau_n) \to 0 \text{ a.s. } (P).$$

Take some positive constant $\beta < 1$. (Later on in the proof we will require that β lie in the interval $(1+\gamma)/2 < \beta < 1$.) Then

$$\text{(A23)} \qquad \frac{V_n^2}{B_n} - \frac{W(\tau_n)^2}{\tau_n} = \frac{W(\tau_n)^2}{\tau_n^{2-\beta}} \frac{\tau_n - \beta_n}{\tau_n^{\beta}} \frac{\tau_n}{B_n}.$$

By the law of the iterated logarithm for Brownian motion (e.g., Shorack and Wellner (1986, p. 27))

$$\limsup_{n\to\infty} \frac{W(\tau_n)}{\{2\tau_n \ln(\ln(\tau_n))\}^{1/2}} = 1,$$

so that

$$\text{(A24)} \qquad W(\tau_n)^2/\tau_n^{2-\beta} \to 0 \text{ a.s. } (P),$$

since $2-\beta > 1$. Next observe that

$$\text{(A25)} \qquad \frac{\tau_n - B_n}{\tau_n^{\beta}} = \frac{\tau_n(1 - B_n/\tau_n)}{\tau_n^{\beta}} \to 0 \text{ a.s. } (P)$$

and $\beta < 1$ imply that $B_n/\tau_n \to 1$ a.s. (P). Hence, in view of (A23) and (A24), it is sufficient for (A22) to prove that (A25) holds. This is easily seen to be equivalent to proving

$$\text{(A26)} \qquad (\tau_n - B_n)/B_n^{\beta} \to 0 \text{ a.s. } (P)$$

for $\beta < 1$, which we now set out to do.

Set $\tau_0 = 0$ and $B_0 = 0$ and define

$$d_j = B_j - B_{j-1} = E(\varepsilon_j^2 | \mathscr{E}_{j-1})$$

and

$$\Delta_j = \tau_j - \tau_{j-1}.$$

Then

$$\tau_n - B_n = \sum_1^n \{(\tau_j - \tau_{j-1}) - B_j - B_{j-1}\} = \sum_1^n (\Delta_j - d_j)$$

and so rewriting (A26) we need to prove

$$\text{(A27)} \qquad B_n^{-\beta} \sum_1^n (\Delta_j - d_j) \to 0 \text{ a.s. } (P).$$

By Kronecker's Lemma, (A27) holds if $\sum_1^\infty (\Delta_j - d_j)/B_j^{\beta} < \infty$, a.s. (P), which holds by Chow's Theorem (Hall-Heyde (1980, p. 35, Theorem 2.17)) if

$$\text{(A28)} \qquad \sum_1^\infty E\left\{[(\Delta_j - d_j)/B_j^{\beta}]^2 | \mathscr{E}_{j-1}\right\} < \infty, \text{ a.s. } (P)$$

since $(\Delta_j - d_j)/B_j^{\beta}$ is a martingale.

Now $E\{(\Delta_j - d_j)^2 | \mathscr{E}_{j-1}\} \le E(\Delta_j^2 | \mathscr{E}_{j-1}) - d_j^2 \le E(\Delta_j^2 | \mathscr{E}_{j-1})$, so that it is sufficient for (A28) to prove that

$$\text{(A29)} \qquad \sum_1^\infty E(\Delta_j^2 | \mathscr{E}_{j-1})/B_j^{2\beta} < \infty \text{ a.s. } (P).$$

Using H^2(ii) we have

$$E(\Delta_j^2 | \mathscr{E}_{j-1}) \le C_2 E(\varepsilon_j^4 | \mathscr{E}_{j-1}) \text{ a.s. } (P)$$

$$\le C_2 C_a \{E(\varepsilon_j^2 | \mathscr{E}_{j-1})\}^2 \text{ a.s. } (P)$$

because of (D9). Therefore, (A29) holds if

$$\text{(A30)} \qquad \sum_{1}^{\infty} (E(\varepsilon_j^2|\mathscr{E}_{j-1}))^2/B_j^{2\beta} < \infty \text{ a.s. } (P)$$

holds. Since $d_j = E(\varepsilon_j^2|\mathscr{E}_{j-1}) = B_j - B_{j-1}$, we may write the left side of (A30) as

$$\text{(A31)} \qquad \sum_{1}^{\infty} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) E(\varepsilon_j^2|\mathscr{E}_{j-1}).$$

Now take some $M > 0$, possibly large. Then by (D10) we have

$$\text{(A32)} \qquad P[E(\varepsilon_n^2|\mathscr{E}_{n-1})/B_n^{\gamma} > M \text{ at most finitely often}] = 1.$$

The event

$$\text{(A33)} \qquad [E(\varepsilon_n^2|\mathscr{E}_{n-1})/B_n^{\gamma} > M \text{ at most finitely often}]$$

implies the event

$$\left[\sum_{1}^{\infty} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) E(\varepsilon_j^2|\mathscr{E}_{j-1}) \le \sum_{N+1}^{\infty} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) MB_j^{\gamma} + \sum_{1}^{N} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) E(\varepsilon_j^2|\mathscr{E}_{j-1}) \text{ for some finite } N\right]$$

which implies

$$\text{(A34)} \qquad \left[\sum_{1}^{\infty} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) E(\varepsilon_j^2|\mathscr{E}_{j-1}) \le M \sum_{1}^{\infty} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta-\gamma}}\right) + \sum_{1}^{N} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) E(\varepsilon_j^2|\mathscr{E}_{j-1}) \text{ for some finite } N\right].$$

Let $p = 2\beta - \gamma$ and since (D10) holds for γ with $0 < \gamma < 1$ we may choose β in the interval $(1 + \gamma)/2 < \beta < 1$ and then $p = 2\beta - \gamma > 1$. We have

$$\sum_{1}^{\infty} (B_j - B_{j-1})/B_j^p = \sum_{1}^{\infty} d_j/B_j^p,$$

where $B_j = B_{j-1} + d_j = \sum_1^j d_k$. Since $d_k \ge 0$ a.s. (P) for all k and $B_j \to \infty$ a.s. (P) as $j \to \infty$ by (D2), it follows by Dini's Theorem (e.g., Knopp (1956, Theorem 1, p. 125)) that

$$\text{(A35)} \qquad \sum_{1}^{\infty} d_j/B_j^p < \infty \text{ a.s. } (P)$$

because $p > 1$.

Event (A33) implies (A34) which, because of (A35), implies

$$\text{(A36)} \qquad \left[\sum_{1}^{\infty} \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) E(\varepsilon_j^2|\mathscr{E}_{j-1}) < \infty\right].$$

In view of (A32) we deduce that

$$P\left[\sum_1^\infty \left(\frac{B_j - B_{j-1}}{B_j^{2\beta}}\right) E(\varepsilon_j^2 | \mathscr{E}_{j-1}) < \infty\right] \geq P[E(\varepsilon_n^2 | \mathscr{E}_{n-1}) B_n^\gamma > M \text{ at most finitely often}] = 1$$

thereby proving (A30). This in turn establishes (A27), (A26), and thus (A22), which gives the stated result (30).

PROOF OF THEOREM 4.1: The proof of Theorem 2.1 carries over almost verbatim. We need to replace the quadratic variation A_t with the conditional variance matrix process $B_n(\theta) = \langle l_n^{(1)}(\theta) \rangle$, and the univariate Taylor series expansion (A5) is replaced by the corresponding multivariate expansion. Writing $\theta - \hat{\theta}_n = \lambda h$ with $h \in S_k$ we have the equality

$$\begin{aligned}(\text{A6}') \qquad (\theta - \hat{\theta}_n) l_n^{(2)}(\theta_m)(\theta - \hat{\theta}_n) = &-(\theta - \hat{\theta}_n)' B_n(\theta - \hat{\theta}_n) + \{h'[l_n^{(2)}(\theta_m) - l_n^{(2)}(\theta^0)]h/h'B_n h \\ &+ h'[l_n^{(2)}(\theta^0) + B_n]h/h'B_n h\}(\theta - \hat{\theta}_n)' B_n(\theta - \hat{\theta}_n)\end{aligned}$$

in place of (A6). Then, using (D3′) and (D4′) we obtain

$$(\text{A9}') \qquad I_1 = |B_n|^{-1/2} \exp\{l_n(\hat{\theta}_n)\}(2\pi)^{k/2} \pi_0[1 + o_{as}(1)]$$

in place of (A9). The rest of the proof proceeds as before and we obtain

$$d\mathscr{P}_n / dP_t^0 = c_0 |B_n|^{-1/2} \exp\{l_n(\hat{\theta}_n)\}[1 + o_{as}(1)]$$

giving (38). The other two representations (36) and (37) follow in the same way as the univariate case.

PROOF OF THEOREM 4.6: The PIC statistic is

$$dQ_n / dP_n(\hat{\sigma}^2) = (\Sigma_1^n Y_{t-1}^2 / \hat{\sigma}^2)^{-1/2} \exp\{(1/2)\hat{h}_n^2 \Sigma_1^n Y_{t-1}^2 / \hat{\sigma}^2\}$$

where $\hat{\sigma}^2 = n^{-1} \Sigma_1^n (\Delta Y_t - \hat{h}_n Y_{t-1})^2$. When there is a unit root in (41′), i.e. when $h = 0$, we have

$$\hat{h}_n^2 \sum_1^n Y_{t-1}^2 / \hat{\sigma}^2 = [n(\hat{\alpha}_n - 1)]^2 \left[n^{-2} \sum_1^n Y_{t-1}^2 / \hat{\sigma}^2\right] = O_p(1) \quad \text{and} \quad \sum_1^n Y_{t-1}^2 / \hat{\sigma}^2 = O_p(n^2)$$

as $n \to \infty$, so that

$$dQ_n / dP_n(\hat{\sigma}^2) \to_p 0.$$

Hence, $P(dQ_n / dP_n(\hat{\sigma}^2) > 1) \to 0$ as $n \to \infty$ and the type I error tends to zero as $n \to \infty$.

When the model does not have a unit root ($h \neq 0$) and $|\alpha| < 1$, say, then

$$\hat{h}_n^2 \sum_1^n Y_{t-1}^2 / \hat{\sigma}^2 = [\sqrt{n}(\hat{\alpha}_n - 1)]^2 \left[n^{-1} \sum_1^n Y_{t-1}^2 / \hat{\sigma}^2\right] = O_p(n)$$

and $\Sigma_1^n Y_{t-1}^2 / \hat{\sigma}^2 = O_p(n)$, so that

$$\ln[dQ_n / dP_n(\hat{\sigma}^2)] = O_p(n)$$

as $n \to \infty$. It follows that $dQ_n / dP_n(\hat{\sigma}^2)$ diverges as $n \to \infty$ and $P_n^h(dQ_n / dP_n(\hat{\sigma}^2) > 1) \to 1$ as $n \to \infty$. Thus, the power of the test tends to unity and the type II error tends to zero as $n \to \infty$. By a similar argument the same behavior obtains when $\alpha > 1$.

PROOF OF LEMMA B: Note that $A_n = A_{n-1} + Y_{n-1}^2 = A_{n-1}(1 + Y_{n-1}^2 / A_{n-1})$ and thus by recursion we have:

$$(\text{A37}) \qquad A_n = A_{n_0} \prod_{i=0}^{n-n_0-1} (1 + Y_{n_0+i}^2 / A_{n_0+i}) = A_{n_0} \prod_{t=n_0+1}^{n} g_t, \quad \text{with } g_t = 1 + Y_t^2 / A_t = f_t / \sigma^2.$$

Next

$$V_n^2 A_n^{-1} - V_{n-1}^2 A_{n-1}^{-1} = A_n^{-1}\{(V_{n-1} + Y_{n-1}\Delta Y_n)^2 - V_{n-1}^2(1 + Y_{n-1}^2/A_{n-1})\}$$
$$= A_n^{-1}\{2\hat{h}_{n-1}A_{n-1}Y_{n-1}\Delta Y_n + (\Delta Y_n Y_{n-1})^2 - \hat{h}_{n-1}^2 A_{n-1}Y_{n-1}^2\}$$
$$= A_n^{-1}\left\{-(\Delta Y_n - \hat{h}_{n-1}Y_{n-1})^2 A_{n-1} + (\Delta Y_n)^2 A_n\right\}$$
$$= -(\Delta Y_n - \hat{h}_{n-1}Y_{n-1})^2(A_{n-1}/A_n) + (\Delta Y_n)^2$$
$$= -(\Delta Y_n - \hat{h}_{n-1}Y_{n-1})^2/g_n + (\Delta Y_n)^2$$

and by recursion we have

$$V_n^2 A_n^{-1} - V_{n_0}^2 A_{n_0}^{-1} = -\sum_{t=n_0+1}^{n} (\Delta Y_t - \hat{h}_{t-1}Y_{t-1})^2/g_t + \sum_{t=n_0+1}^{n} (\Delta Y_t)^2. \tag{A38}$$

Combining (A37) and (A38) in (46) we get

$$r_n(\sigma^2) = (A_n/A_{n_0})^{-1/2}\exp\{(1/2\sigma^2)V_n^2 A_n^{-1} - (1/2\sigma^2)V_{n_0}^2 A_{n_0}^{-1}\}$$
$$= \prod_{t=n_0+1}^{n} \frac{(2\pi f_t)^{-1/2}\exp\left\{-(1/2f_t)(\Delta Y_t - \hat{h}_{t-1}Y_{t-1})^2\right\}}{(1/2\pi\sigma^2)^{1/2}\exp\{-(1/2\sigma^2)(\Delta Y_t)^2\}}$$

as required.

REFERENCES

BERGER, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis* (2nd Edition). New York: Springer Verlag.

BROWN, R. L., J. DURBIN, AND J. M. EVANS (1975): "Techniques for Testing the Constancy of Regression Relationships over Time" (with discussion), *Journal of the Royal Statistical Society*, Series B, 37, 149–192.

CHAO, J., AND P. C. B. PHILLIPS (1994): "Bayesian Model Selection in Partially Nonstationary Vector Autoregressive Processes with Reduced Rank Regression Structure," mimeographed, Yale University.

DONSKER, M. D., AND S. R. S. VARADHAN (1977): "On Laws of the Iterated Logarithm for Local Times," *Communications in Pure and Applied Mathematics*, 30, 707–753.

Econometric Theory (1994): *Yale-NSF Symposium on Bayes Methods and Unit Roots*, Vol. 10, No. 3/4.

EMERY, M. (1989): *Stochastic Calculus in Manifolds*. New York: Springer Verlag.

HALL, P., AND C. C. HEYDE (1980): *Martingale Limit Theory and its Application*. New York: Academic Press.

HARTIGAN, J. A. (1983): *Bayes Theory*. New York: Springer Verlag.

IBRAGIMOV, I. A., AND R. Z. HAS'MINSKII (1981): *Statistical Estimation: Asymptotic Theory*. New York: Springer Verlag.

IKEDA, N., AND S. WATANABE (1989): *Stochastic Differential Equations and Diffusion Processes* (Second Edition). Amsterdam: North Holland.

Journal of Applied Econometrics (1991): *Classical and Bayesian Methods of Testing for Unit Roots*, themed issue.

KARATZAS, I., AND S. E. SHREVE (1991): *Brownian Motion and Stochastic Calculus* (2nd Ed.). New York: Springer Verlag.

KNOPP, K. (1956): *Infinite Sequences and Series*. New York: Dover.

LAI, T. L., AND C. Z. WEI (1983): "Asymptotic Properties of General Autoregression Models and Strong Consistency of Least Squares Estimates to their Parameters," *Journal of Multivariate Analysis*, 12, 346–370.

LEAMER, E. E. (1978): *Specification Searches: Ad hoc Inferences with Nonexperimental Data*. New York: John Wiley & Sons.

MEYER, P. A. (1989): "A Short Presentation of Stochastic Calculus," in *Stochastic Calculus in Manifolds*, ed. by M. Emery. New York: Springer Verlag.

PHILLIPS, P. C. B. (1987): "Time Series Regression with a Unit Root," *Econometrica*, 55, 277–301.

——— (1991): "Optimal Inference in Cointegrated Systems," *Econometrica*, 59, 283–306.

——— (1994): "Model Determination and Macroeconomic Activity," Cowles Foundation Discussion Paper No. 1083, to appear in *Econometrica*, 1996.

PHILLIPS, P. C. B., AND W. PLOBERGER (1994): "Posterior Odds Testing for a Unit Root with Data-based Model Selection," *Econometric Theory*, 10, 774–808.

PROTTER, P. (1991): *Stochastic Integration and Differential Equations: A New Approach*. New York: Springer Verlag.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

SHORACK, G. R., AND J. A. WELLER (1986): *Empirical Processes with Applications to Statistics*. New York: Wiley.

STRASSER, H. (1986): "Martingale Difference Arrays and Stochastic Integrals," *Probability Theory and Related Fields*, 72, 83–98.

WALKER, A. M. (1969): "Asymptotic Behavior of Posterior Distributions," *Journal of the Royal Statistical Society*, Series B, 31, 80–88.

ZELLNER, A. (1978): "Jeffreys-Bayes Posterior Odds Ratio and the Akaike Information Criterion for Discriminating between Models," *Economics Letters*, 1, 337–342.