COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale University
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1017

POSTERIOR ODDS TESTING FOR A UNIT ROOT WITH DATA-BASED MODEL SELECTION

by

Peter C. B. Phillips and Werner Ploberger

May 1992

POSTERIOR ODDS TESTING FOR A UNIT ROOT WITH DATA-BASED MODEL SELECTION*

by

Peter C. B. Phillips
Cowles Foundation for Research in Economics
Yale University

and

Werner Ploberger
Institute for Econometrics and Operations Research
Technical University of Vienna

0. ABSTRACT

The Kalman filter is used to derive updating equations for the Bayesian data density in discrete time linear regression models with stochastic regressors. The implied "Bayes model" has time varying parameters and conditionally heterogeneous error variances. A σ-finite "Bayes model" measure is given and used to produce a new model selection criterion (PIC) and objective posterior odds tests for sharp null hypotheses like the presence of a unit root. This extends earlier work in Phillips-Ploberger (1991). Autoregressive-moving average (ARMA) models are considered and a general test of trend stationarity versus difference stationarity is developed in ARMA models allowing for automatic order selection of the stochastic regressors and the degree of the deterministic trend. The tests are completely consistent in that both type I and type II errors tend to zero as the sample size tends to infinity.

Simulation results and an empirical application are reported. The simulations show that the new model selection criterion "PIC" works very well and is generally superior to the Schwarz criterion BIC even in stationary systems. Empirical application of our methods to the Nelson-Plosser (1982) series show that three series (unemployment, industrial production and the money stock) are level or trend stationary. The other eleven series are found to be stochastically nonstationary.

March, 1992

*All of the computations reported in this paper were performed by the first author on a 486-33 PC using programs written in GAUSS-386i (Version 2.2). The paper was written while Werner Ploberger was a visitor at the Cowles Foundation during 1991-1992. Phillips thanks the NSF for research support under Grant No. SES 8821180. Ploberger thanks the Fonds zur Förderung der wissenschaftlichen Forschung for supporting his stay at Yale with Schrödingerstipendium Nr. J0469-TEC. The authors thank Glena Ames for her skill and effort in keyboarding the manuscript.

1. INTRODUCTION

Reasons for the apparent divergence of classical and Bayesian statistical inference in time series applications have been explored in very recent work by the authors that gives attention to the role of both prior distributions and data conditioning in the practical implementation of Bayesian methods. The impact of prior distributions on Bayesian inference with time dependent data was considered in Phillips (1991), which addressed the issue of determining "impartial" or "objective" priors for the parameters in simple time series models and showed the sensitivity of Bayesian posteriors to both priors and model specification in that context. The role of data conditioning in Bayesian analysis of time series was studied in Phillips-Ploberger (1991). The operation of the likelihood principle that underlies Bayesian inference ensures that Bayesian time series analysis is conducted conditional on the realized history of the process. The Phillips-Ploberger paper examines the first order autoregressive model (AR(1)) and shows that the mathematical effect of this data conditioning in inference is to translate the model (and its reference probability measure) to what we call in that paper a "Bayes model" (and, respectively, "Bayes model" measure) in which the parameters are time varying and data dependent. Methodologically, the Bayesian approach involves no commitment to any "true" value of a parameter (unlike the classical approach). But, given a particular historical trajectory, we show in Phillips-Ploberger (1991) that the use of the likelihood principle in fact commits the investigator to a new model in which the parameters evolve according to the latest best estimate from the data available to that point on the trajectory. Phillips-Ploberger use this conceptual framework to construct a new test of one "Bayes model" against another. The test is a special type of posterior odds test and is based on the Radon Nikodym (RN) derivative of the respective "Bayes model" measures of the two models. The test can be used to test point null hypotheses (like that of a unit root), it has good finite sample performance and it has interesting asymptotic properties because both type I and type II errors tend to zero as the sample size tends to infinity.

The main purpose of the present paper is to extend the Phillips-Ploberger analysis to a general class of linear discrete time series models that includes ARMA(p, q) models with deterministic trends. Recursive least squares (or Kalman filter) methods are used to derive updating equations for the Bayesian data density. These equations determine the precise form of the "Bayes model" and "Bayes model" probability measure for this general class of time series models. The "Bayes model" measure is used to produce a new model selection criterion (which we term "PIC") that picks the model with the highest posterior density as given by the RN derivative of the "Bayes model" measure of that model with respect to a general "reference model" measure in the class of competing models. This new criterion is, in fact, a generalization of the BIC criterion due to Schwarz (1978) and, indeed, is asymptotically equivalent to BIC in stationary time series models. The PIC criterion is used to select both lag order and deterministic trend degree in the class of ARMA(p, q) models with deterministic trends. We show how to apply this procedure in the context of a recursion that is based on the one suggested originally by Durbin (1962) and Hannan-Rissanen (1982) for the consistent estimation of ARMA models. Following model selection, the posterior odds (PIC) criterion is used again to compare the selected "Bayes model" against the same model with a unit autoregressive root. The procedure provides a completely consistent test for the presence of a unit root in this general class of discrete time series models and gives an algorithm that leads to a data-coherent, parsimonious model choice within this class.

The paper is organized as follows. Section 2 studies a general class of linear discrete time models, derives the respective "Bayes models" and "Bayes model" measures in this class, and explores the martingale structure of these measures. Section 3 develops our new model selection criterion "PIC", gives its asymptotic properties, and shows how it can be used for model selection and for testing point null hypotheses like that of a unit root. Section 4 gives a general application of our methodology to the problem of testing difference versus trend stationarity and provides an algorithm that incorporates both model selection principles and tests of point null hypotheses. Programs for the implementation of this methodology have now been written in GAUSS-386i and Section 5 reports some simple simulation exercises that illus-

trate the performance of the procedure in determining the presence of a unit autoregressive root in models that include AR(p) and ARMA(p, q) models with and without deterministic trends. Overall, these results are considered by the authors to be very encouraging. Section 6 reports an empirical application of our methodology to the Nelson-Plosser data set. The empirical results are striking. Only two series (industrial production and money) are found to have a deterministic trend, only one series (unemployment) is stationary and the remaining eleven series are found to be stochastically nonstationary.

2. "BAYES MODELS" AND "BAYES MODEL" MEASURES IN DISCRETE TIME

The model we consider is a linear regression

$$(1) \qquad y_t = \beta' x_t + \varepsilon_t, \quad (t = 1, 2, \dots)$$

whose dependent variable y_t and error ε_t are real valued stochastic processes on a probability space $(\Omega, \mathcal{F}, P)$. Accompanying y_t is a filtration $\mathcal{F}_t \subset \mathcal{F}$ $(t = 0, 1, 2, \dots)$ to which both y_t and ε_t are adapted. The regressors x_t $(k \times 1)$ in (1) are defined on the same space and are assumed to have the property that x_t is $\mathcal{F}_{t-1}$-measurable. The standard example of (1) will be the autoregression (with $k = p$ lags) given by

$$(2) \qquad y_t = \Sigma_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t,$$

which it will often be more convenient to write as

$$(3) \qquad \Delta y_t = h y_{t-1} + \Sigma_{i=1}^{p-1} \varphi_i \Delta y_{t-i} + \varepsilon_t.$$

We will also consider an augmented version of model (3) with intercept and trend, viz.

$$(4) \qquad \Delta y_t = h y_{t-1} + \Sigma_{i=1}^{p-1} \varphi_i \Delta y_{t-i} + \mu + \gamma t + \varepsilon_t.$$

ARMA models also fit into the general framework of (1), although in this case the regressors are not all observable. Extending (4) by the inclusion of moving average errors we have (with $k = p+q+2$)

$$(5) \qquad \Delta y_t = h y_{t-1} + \Sigma_{i=1}^{p-1} \varphi_i \Delta y_{t-i} + \Sigma_{j=1}^{q} \psi_j \varepsilon_{t-j} + \mu + \gamma t + \varepsilon_t$$

giving an ARMA(p, q) process with trend. In all these examples we will suppose that $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$. In (4) and (5) the parameterization accommodates a unit autoregressive root when $h = 0$. This parameterization is especially convenient when testing for the presence of a unit root. It is also useful in setting up Bayes model alternatives to models like (4) and (5) with a unit root. We shall be explicit about such possibilities later in our discussion.

We introduce the regression notation $Y_n' = [y_1, ..., y_n]$, $X_n' = [x_1, ..., x_n]$, and set $A_n = X_n'X_n$. Let ε_t = iid $N(0, \sigma^2)$ and let us assume for the time being that the error variance σ^2 is known. Then the inference problem presented by (1) is linear in parameters. This facilitates an exact development of our theory and the necessary extensions for σ^2 unknown will be given after this development.

Conditional on $\mathcal{F}_0$ and β the joint density of Y_n with respect to Lebesgue measure (ν) is

$$\text{(6)} \qquad \begin{aligned} \text{pdf}(Y_n|\mathcal{F}_0, \beta) &= dP_n^\beta/d\nu = (2\pi\sigma^2)^{-n/2}\exp\{-(1/2\sigma^2)\textstyle\sum_1^n(y_t - \beta' x_t)^2\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\{-(1/2\sigma^2)[\hat{u}_n'\hat{u}_n + (\hat{\beta}_n - \beta)'A_n(\hat{\beta}_n - \beta)]\} \end{aligned}$$

where $\hat{U}_n = Y_n - X_n\hat{\beta}_n$, $\hat{\beta}_n = (X_n'X_n)^{-1}X_n'Y_n$ and P_n^β is the probability measure of Y_n. The corresponding measure when $\beta = 0$ will be denoted by P_n and will serve as a possible reference measure in the analysis that follows. The likelihood ratio process is then the ratio of these densities, i.e.

$$\text{(7)} \qquad \begin{aligned} L_n(\beta) &= dP_n^\beta/dP_n = \exp\{-(1/2\sigma^2)[-2\beta' X_n'Y_n + \beta' A_n\beta]\} \\ &= \exp\{(1/2\sigma^2)[\hat{\beta}_n'A_n\hat{\beta}_n - (\hat{\beta}_n - \beta)'A_n(\hat{\beta}_n - \beta)]\} . \end{aligned}$$

Combining (6) with a prior density $\pi(\beta)$ for β we have the joint density of (β, Y_n) conditional on $\mathcal{F}_0$, i.e.

$$\text{(8)} \qquad \begin{aligned} \text{pdf}(\beta, Y_n|\mathcal{F}_0) &= \pi(\beta)\text{pdf}(Y_n|\mathcal{F}_0, \beta) \\ &= \left[(2\pi\sigma^2)^{(n-k)/2}|A_n|^{-1/2}\exp\left\{-(1/2\sigma^2)\hat{U}_n'\hat{U}_n\right\}\right] \\ &\quad \times \left[\pi(\beta)(2\pi\sigma^2)^{-k/2}|A_n|^{1/2}\exp\left\{-(1/2\sigma^2)(\hat{\beta}_n - \beta)'A_n(\hat{\beta}_n - \beta)\right\}\right] . \end{aligned}$$

For $\pi(\beta) = \pi_0$ = constant, this expression gives a marginal posterior density process for β of the form

(9) $\Pi_n(\beta) = (2\pi\sigma^2)^{-k/2}|A_n|^{1/2}\exp\left\{-(1/2\sigma^2)(\hat{\beta}_n-\beta)'A_n(\hat{\beta}_n-\beta)\right\} = N(\hat{\beta}_n,\ \sigma^2A_n^{-1})$.

This is the usual Gaussian posterior density for the parameters in the linear regression model with known error variance. The density is centered on the maximum likelihood estimator $\hat{\beta}_n$, based on the data Y_n, and has variance matrix $\sigma^2A_n^{-1}$.

If we integrate β in (8) we obtain the data density for Y_n, viz.

(10) $\text{pdf}(Y_n|\mathcal{F}_0) = \pi_0(2\pi\sigma^2)^{-(n-k)/2}|A_n|^{-1/2}\exp\left\{-(1/2\sigma^2)\hat{U}_n'\hat{U}_n\right\}$.

Let Q_n be the (probability) measure whose density with respect to ν is (10). Then

(11) $Q_n = \int_{\mathbb{R}^k}\pi(\beta)P_n^\beta d\beta$

and thus

$$\begin{aligned}\frac{dQ_n}{dP_n} &= \int_{\mathbb{R}^k}\pi(\beta)\frac{dP_n^\beta}{dP_n}d\beta \\ &= \int_{\mathbb{R}^k}\pi(\beta)L_n(\beta)d\beta \\ &= \pi_0(2\pi\sigma^2)^{k/2}|A_n|^{-1/2}\exp\left\{(1/2\sigma^2)\hat{\beta}_n'A_n\hat{\beta}_n\right\} .\end{aligned}$$

Next observe that (taking expectation with respect to the reference measure P_n)

$$E\left[\frac{dQ_n}{dP_n}|\mathcal{F}_{n-1}\right] = \int_{\mathbb{R}^k}\pi(\beta)E(L_n(\beta)|\mathcal{F}_{n-1})d\beta .$$

Using (7), we can compute the conditional expectation of the likelihood directly as

$$\begin{aligned}E(L_n(\beta)|\mathcal{F}_{n-1}) &= \exp\{-(1/2\sigma^2)(-2\beta'X_{n-1}'Y_{n-1} + \beta'A_n\beta)\}E(\exp\{(1/\sigma^2)\beta'x_ny_n\}|\mathcal{F}_{n-1}) \\ &= \exp\{-(1/2\sigma^2)(-2\beta'X_{n-1}'Y_{n-1} + \beta'A_n\beta)\}\exp\{(1/\sigma^2)\beta'x_nx_n'\beta\} \\ &= \exp\{-(1/2\sigma^2)(-2\beta'X_{n-1}'Y_{n-1} + \beta'A_{n-1}\beta)\} \\ &= L_{n-1}(\beta) ,\end{aligned}$$

using the fact that $y_n|\mathcal{F}_{n-1} = N(0,\ \sigma^2)$ under P_n. Hence,

$$E\left[\frac{dQ_n}{dP_n}|\mathcal{F}_{n-1}\right] = \int_{\mathbb{R}^k}\pi(\beta)L_{n-1}(\beta)d\beta = \frac{dQ_{n-1}}{dP_{n-1}} ,$$

and dQ_n/dP_n satisfies the martingale property under the probability measure P_n. Notice that this conditional expectation is finite even though $E(dQ_n/dP_n)$ is not finite (and dQ_n/dP_n is therefore not integrable) when $\pi(\beta) = \pi_0$ is constant.

Collecting these results together, we have:

2.1. THEOREM: *Under the uniform prior* $\pi(\beta) = \pi_0 = (2\pi)^{-k/2}$, *the (probability) density of the data* $Y_n = [y_1, ..., y_n]$ *generated by the model* (1), *conditional on* $\mathcal{F}_0$, *and taken with respect to the reference measure* P_n *is given by*

$$(12) \qquad \frac{dQ_n}{dP_n} = \left|(1/\sigma^2)A_n\right|^{-1/2} \exp\left\{(1/2\sigma^2)\hat{\beta}_n' A_n \hat{\beta}_n\right\} .$$

The density process dQ_n/dP_n *is a local* P_n*-martingale. Although* dQ_n/dP_n *is not integrable it has finite conditional expectation and satisfies the martingale property*

$$E\left[\frac{dQ_n}{dP_n}\middle|\mathcal{F}_{n-1}\right] = \frac{dQ_{n-1}}{dP_{n-1}} \quad \text{a.s.}$$

under the reference measure P_n. □

2.2. REMARKS

(i) Expression (12) is the likelihood ratio of the measure Q_n with respect to the base measure P_n. It may be used for both hypothesis testing and model selection purposes, as we will explain below.

(ii) The density $dQ_n/d\nu = \text{pdf}(Y_n|\mathcal{F}_0)$ given by (10) is not integrable over the space $\mathbb{R}^n$ which supports the variate matrix Y_n and thus Q_n is a σ-finite measure rather than a proper probability measure on $\mathbb{R}^n$. Hence, the use of the parentheses around the word "probability" in the statement of the theorem. This feature of Q_n is the consequence of the use of an improper prior $\pi(\beta) = \pi_0$ on β over $\mathbb{R}^k$. As we shall show below, the conditional measures based on Q_n are, in contrast to Q_n itself, proper probability measures and it is the sequence of conditional measures that defines the characteristics of the Bayesian solution to the problem of inference in the model (1). We shall use these conditional measures to construct a model,

which we call the "Bayes model" of the data, that Bayesian inference implicitly uses in place of (1).

(iii) The general form of the RN derivative dQ_n/dP_n that is given in (12) is invariant in large samples to the use of a wide class of continuous prior densities $\pi(\ )$. To see this, we note that if the excitation condition (i.e. $\lambda_{\min}(A_n) \to \infty$ as $n \to \infty$) holds we may apply the Laplace approximation to the integral defining dQ_n/dP_n giving

$$\frac{dQ_n}{dP_n} = \int_{\mathbf{R}^k} \pi(\beta) L_n(\beta) d\beta \sim \pi(\hat{\beta}_n)(2\pi\sigma^2)^{k/2} |A_n|^{-1/2} \exp\left\{(1/2\sigma^2)\hat{\beta}_n' A_n \hat{\beta}_n\right\} .$$

Since $\pi(\hat{\beta}_n) \to_{a.s.} \pi(\beta)$ as $n \to \infty$ when the excitation condition applies, dQ_n/dP_n is asymptotically proportional to the formula given in (12). This argument continues to hold in quite general nonlinear models, leading to the same general formula (12). A complete treatment of the general case is presently being written up and will be reported in a later paper (see Phillips-Ploberger, 1992a). □

First, it is of interest to define the model of the data for which Q_n is the (probability) measure. From (11) it is apparent that this measure is actually a weighted average of the measures P_n^β (whose density $dP_n^\beta/d\nu$ is given in (6)) with weights delivered by the prior density $\pi(\beta)$. The measure Q_n is, as we have remarked above, a σ-finite measure for the data Y_n. Note that (12) becomes undefined when there is insufficient data (i.e., $n < k$) to determine $\hat{\beta}_n$ and in such cases $|A_n| = |X_n'X_n| = 0$. It is therefore appropriate to regard (12) as defining an admissible measure for the data provided a minimal amount of information has already accumulated so that $n \geq k$ and $|A_n| = |X_n'X_n| > 0$. As we shall see, conditional on the accumulation of such minimal information to initialize the process, Q_n as given by (12) leads to a proper conditional probability measure and this conditional measure defines a new (Bayesian) model for the data that replaces the classical model (1).

Consider the data density $\text{pdf}(Y_n|\mathcal{F}_0)$ given by (10). An alternative way of writing this density is to employ the prediction error decomposition, viz.

$$
\begin{aligned}
\text{pdf}(Y_n|\mathcal{F}_0) &= \prod_{t=k+1}^{n} \text{pdf}(y_t|\mathcal{F}_{t-1})\text{pdf}(Y_k|\mathcal{F}_0) \\
&= (2\pi)^{-(n-k)/2}\prod_{t=k+1}^{n}\left[f_t^{-1/2}\exp\left\{-(1/2f_t)v_t^2\right\}\right]\text{pdf}(Y_k|\mathcal{F}_0) \ .
\end{aligned} \tag{13}
$$

In this decomposition

$$v_t = y_t - \hat{y}_{t|t-1} = y_t - \hat{\beta}_{t-1}'x_t \ , \ t = k+1, \ ..., \ n$$

are the prediction errors and $\hat{y}_{t|t-1}$ is the best forecast of y_t using information available up to time $t-1$, i.e. information in $\mathcal{F}_{t-1}$. The forecast error variance conditional on $\mathcal{F}_{t-1}$ in (13) is given by

$$f_t = \sigma^2(1 + x_t'A_{t-1}^{-1}x_t), \ t = k+1, \ ..., \ n$$

and the conditional distribution of v_t given $\mathcal{F}_{t-1}$ is

$$v_t|_{\mathcal{F}_{t-1}} = N(0, f_t) \ .$$

Expression (13) is simply derived from the formulae for recursive least squares (see Brown, Durbin and Evans (1975)). Note that

$$
\begin{aligned}
|A_n| &= |A_{n-1} + x_nx_n'| = |A_{n-1}|(1 + x_n'A_{n-1}^{-1}x_n) \\
&= \left[\prod_{t=k+1}^{n}(1 + x_t'A_{t-1}^{-1}x_t)\right]|A_k| \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{U}_n'\hat{U}_n &= (Y_n - X_n\hat{\beta}_n)'(Y_n - X_n\hat{\beta}_n) \\
&= \hat{U}_{n-1}'\hat{U}_{n-1} + (y_n - x_n'\hat{\beta}_{n-1})^2\left\{1 - x_n'A_n^{-1}x_n\right\} \\
&= \hat{U}_{n-1}'\hat{U}_{n-1} + (y_n - x_n'\hat{\beta}_{n-1})^2/(f_n/\sigma^2) \\
&= \hat{U}_k'\hat{U}_k + \Sigma_{t=k+1}^{n}v_t^2/(f_t/\sigma^2) \\
&= \Sigma_{t=k+1}^{n}v_t^2/(f_t/\sigma^2) \ ,
\end{aligned}
$$

since $\hat{U}_k = Y_k - X_k\hat{\beta}_k = 0$. To establish equivalence between (10) and (13) we simply set the initial conditional density of Y_k to be

$$\text{pdf}(Y_k|\mathcal{F}_0) = \pi_0|A_k|^{-1/2} ,$$

which is uniform on $\mathbb{R}^k$.

Next, observe that

$$dQ_n/dv = \text{pdf}(Y_n|\mathcal{F}_0) = \text{pdf}(Y_k|\mathcal{F}_0)(2\pi)^{-(n-k)/2}\prod_{t=k+1}^{n}\left[f_t^{-1/2}\exp\left\{-(1/2f_t)v_t^2\right\}\right] ,$$

so that

$$dQ_n/dQ_{n-1} = (2\pi f_n)^{-1/2}\exp\left\{-(1/2f_n)v_n^2\right\} = N(0, f_n) ,$$

giving the conditional density of the data at the latest observation, n, using information on the trajectory up to the time period $n-1$. This conditional density is proper and holds for all $n > k$, leading to the following general statement of the result.

2.3. THEOREM: *The Bayesian conditional density of the observation y_t given $\mathcal{F}_{t-1}$ (i.e. information on the historical trajectory up to time $t-1$) is*

(14) $\quad dQ_t/dQ_{t-1} = \text{pdf}(Y_t|\mathcal{F}_{t-1}) = (2\pi f_t)^{-1/2}\exp\{-(1/2f_t)v_t^2\} = N(0, f_t)$, $t = k+1, k+2, \ldots$

The "Bayes model" corresponding to this data density is

(15) $\quad y_t = \hat{\beta}'_{t-1}x_t + v_t$, where $v_t|\mathcal{F}_{t-1} = N(0, f_t)$

that is,

(16) $\quad E(v_t|\mathcal{F}_{t-1}) = 0$, $E(v_t^2|\mathcal{F}_{t-1}) = f_t = \sigma^2\{1 + x_t'A_{t-1}^{-1}x_t\}$

and $\hat{\beta}_{t-1} = (X'_{t-1}X_{t-1})^{-1}X'_{t-1}Y_{t-1}$ is the least squares estimate based on information in $\mathcal{F}_{t-1}$. □

2.4. REMARKS

(i) In contrast to (1), the Bayes model (15) is a time varying parameter model where $\hat{\beta}_{t-1}$ evolves according to the best estimate of the slope coefficient that is available from the latest data. Note that the error process in (15) is conditionally heterogeneous with conditional vari-

ance f_t, as given in (16), explicitly dependent on the past data. The form of the conditional heterogeneity is nonlinear in $\{x_s : s = t, t-1, ..., 1\}$. As information about the process y_t accumulates (i.e., as $\lambda_{min}(A_t) \to \infty$), the conditional variance f_t tends to a constant value σ^2. This formulation of the Bayes model for the data extends the one developed in Phillips-Ploberger (1991) for the AR(1) model.

(ii) The prediction error decomposition of the density $\text{pdf}(Y_n|\mathcal{F}_0)$, from which (14) is derived, is closely related to the traditional prediction error formulation of the likelihood function that is based on the Kalman filter (see, e.g. Harvey (1989)). There is, however, a major difference in the use of the Kalman updating algorithm in these two cases. In the traditional use of the updating algorithm, it is the likelihood function (as a function of the parameter vector β) that is extracted. In that case the optimal predictor of y_t from $\mathcal{F}_{t-1}$ would be $y_{t|t-1} = \beta' x_t$ and, in place of (15), we would have

$$y_t = \beta' x_t + v_t \text{ , where } v_t|_{\mathcal{F}_{t-1}} = N(0, \sigma^2) \text{ , } t = k+1, ..., n \text{ .}$$

That is, the Kalman filter produces the exact likelihood (conditional on $\mathcal{F}_k$ *and* β) of the classical statistical model (1) that we started with. By contrast, in our use of the updating algorithm it is the Bayesian data density $\text{pdf}(y_t|\mathcal{F}_{t-1})$ that is generated by the algorithm. It is this data density, given by (14), that prescribes the implied "Bayes model" of the data.

(iii) The data density (14) is exact, as is the "Bayes model" (15). When we relax the Gaussian error assumption and the uniform prior assumption in (1), under both of which the density $\text{pdf}(Y_n|\mathcal{F}_0)$ given in (10) is derived, the data density (14) holds only approximately. The same is true when the error variance σ^2 in (1) is treated as an unknown parameter. In these cases a large sample approximation theory for the likelihood leads to an asymptotic density and "Bayes model" that have the same form as (14) and (15). In the latter case, since it is assumed in the development of the asymptotics that $\lambda_{min}(A_t) \to \infty$ as $t \to \infty$, the conditional variance of the prediction error in the approximate "Bayes model" is $E(v_t^2|\mathcal{F}_{t-1}) = \sigma^2$ rather than (15), which applies exactly in the Gaussian case. A general theory that covers these cases will be reported elsewhere (see Phillips-Ploberger (1992a)).

(iv) The "Bayes model" (15) can be interpreted as a simple "location model" in which $\hat{\beta}_{t-1}' x_t$ provides the best estimate using data from $\mathcal{F}_{t-1}$ of the location of the latest observation y_t. Another way to express this idea is as follows. Bayesian analysis proceeds by conditioning on the observed historical trajectory. As we move along such a trajectory, the best Bayesian estimate (delivered by the posterior mean of the conditional predictive density of y_t) given the data record in $\mathcal{F}_{t-1}$ is $y_{t|t-1} = \hat{\beta}_{t-1}' x_t$. Thus, Bayesian inference about y_t is centered on $\hat{\beta}_{t-1}' x_t$. Under Gaussian assumptions about the errors in the model and a uniform prior on the coefficients we get, in place of the original time series model (1), the simple location model (15) with Gaussian errors v_t. The only complication is that this model evolves period by period and is conditional on the historical record to $\mathcal{F}_{t-1}$.

(v) We call (15) the "Bayes model" because it is the exact model for the data that is implied by the use of traditional Bayes methods under a Gaussian likelihood and (improper) uniform prior. Such methods lead to the Gaussian posterior density $N(\hat{\beta}_n, \sigma^2 A_n^{-1})$ given by (9) above when working with the full sample of data Y_n. This density is obtained by taking the product of the prior and the likelihood (viz. $\pi(\beta)L_n(\beta)$) and by rescaling to achieve a proper density; i.e. the posterior density (9) is

$$(17) \qquad \Pi_n(\beta) = \pi(\beta)L_n(\beta)/\int_{\mathbb{R}^k} \pi(\beta)L_n(\beta)d\beta = \pi(\beta)L_n(\beta)/(dQ_n/dP_n) ,$$

where the last equality follows from (12). Noting that $L_n(\beta) = dP_n^\beta/dP_n$, we deduce that

$$\Pi_n(\beta) = \pi(\beta)dP_n^\beta/dQ_n .$$

Thus, the posterior density $\Pi_n(\beta) = N(\hat{\beta}_n, \sigma^2 A_n^{-1})$ is the direct outcome of employing the likelihood ratio $L_n^Q(\beta) = dP_n^\beta/dQ_n$. This is the density of the measure P_n^β of the model (1) taken with respect to the measure Q_n of data Y_n generated by the time varying parameter model (15). Under this transformation of the measures, the reference measure P_n is replaced by the "Bayes model" measure Q_n in constructing the likelihood. With this new reference measure, associated with model (15), it is natural that inference about β be centered on $\hat{\beta}_n$ when it is conducted through the posterior $\Pi_n(\beta)$. In this respect Bayesian analysis of the time series regression model (1) is identical to Bayesian analysis on the linear regression model with fixed

regressors. The equivalence is the result of data conditioning and the implicit use of the measure Q_n in constructing the likelihood.

2.5. THEOREM

(a) *The least squares estimator* $\hat{\beta}_t$ *is a local* Q_t*-martingale, has finite conditional expectation under the measure* Q_t *and satisfies the martingale property*

$$E_{Q_t}(\hat{\beta}_t|\mathcal{F}_{t-1}) = \hat{\beta}_{t-1} .$$

(b) *Under the Bayes measure* Q_t *the conditional distribution of* $\hat{\beta}_t$ *given* $\mathcal{F}_{t-1}$ *is normal with mean* $\hat{\beta}_{t-1}$ *and covariance matrix* $\sigma^2(A_{t-1}^{-1} - A_t^{-1})$, i.e.

$$\hat{\beta}_t|_{\mathcal{F}_{t-1}} =_d N(\hat{\beta}_{t-1}, \sigma^2(A_{t-1}^{-1} - A_t^{-1})) = N(\hat{\beta}_{t-1}, f_t A_t^{-1} x_t x_t' A_t^{-1}) .$$

(c) *Under the Bayes measure* Q_n*, the posterior distribution* $\Pi_t = N(\hat{\beta}_n, \sigma^2 A_t^{-1})$ *is a local martingale, has finite conditional expectation and satisfies the martingale property*

$$E_{Q_t}[N(\hat{\beta}_n, \sigma^2 A_t^{-1})|\mathcal{F}_{t-1}] = N(\hat{\beta}_{t-1}, \sigma^2 A_{t-1}^{-1}) .$$

2.6. PROOF OF THEOREM 2.5

(a) From recursive least squares formulae (e.g. Brown-Durbin-Evans, 1976, lemma 2) we have

$$\hat{\beta}_t = \hat{\beta}_{t-1} + (X_t' X_t)^{-1} x_t (y_t - x_t' \hat{\beta}_{t-1}) = \hat{\beta}_{t-1} + A_t^{-1} x_t v_t .$$

Under Q_t we have $E_{Q_t}(\hat{\beta}_t|\mathcal{F}_{t-1}) = 0$ by Theorem 2.3 and hence

$$E_{Q_t}(\hat{\beta}_t|\mathcal{F}_{t-1}) = \hat{\beta}_{t-1} ,$$

as required. The process $\hat{\beta}_t$ is a local Q_t-martingale because Q_t is σ-finite and thus $\hat{\beta}_t$ is not Q_t integrable, i.e. $E_{Q_t}(\hat{\beta}_t)$ does not exist.

(b) From Theorem 2.3 we have $v_t|_{\mathcal{F}_{t-1}} = N(0, f_t)$ and hence

$$\hat{\beta}_t|\mathcal{F}_{t-1} =_d N(\hat{\beta}_{t-1}, f_t A_t^{-1} x_t x_t' A_t^{-1}) .$$

Next, observe that

$$A_t^{-1} = (A_{t-1} + x_t x_t')^{-1} = A_{t-1}^{-1} - A_{t-1}^{-1} x_t x_t' A_{t-1}^{-1}/g_t ,$$

where $g_t = 1 + x_t' A_{t-1}^{-1} x_t$. Also

$$A_{t-1}^{-1} = (A_t - x_t x_t')^{-1} = A_t^{-1} + A_t^{-1} x_t x_t' A_t^{-1}/h_t$$

where $h_t = 1 - x_t' A_t^{-1} x_t$. Now

$$\begin{aligned} h_t &= 1 - x_t' A_{t-1} x_t + (x_t' A_{t-1} x_t)^2/g_t \\ &= (1/g_t)\{g_t - x_t' A_{t-1} x_t g_t + (x_t' A_{t-1} x_t)^2\} \\ &= 1/g_t . \end{aligned}$$

Hence,

$$A_{t-1}^{-1} - A_t^{-1} = g_t A_t^{-1} x_t x_t' A_t^{-1}$$

and since $f_t = \sigma^2 g_t$ we deduce that

$$\hat{\beta}_t|_{\mathcal{F}_{t-1}} =_d N(\hat{\beta}_{t-1}, \sigma^2(A_{t-1}^{-1} - A_t^{-1}))$$

as required.

(c) The characteristic function of the posterior distribution Π, is

$$cf_t(s) = \exp\{is'\hat{\beta}_t - (1/2)\sigma^2 s' A_t^{-1} s\} .$$

Now

$$\begin{aligned} E_{Q_t}(cf_t(s)|\mathcal{F}_{t-1}) &= \exp\{is'\hat{\beta}_{t-1} - (1/2)\sigma^2 s' A_t^{-1} s\} E_{Q_t}[\exp(is' A_t^{-1} x_t v_t)|\mathcal{F}_{t-1}] \\ &= \exp\{is'\hat{\beta}_{t-1} - (1/2)\sigma^2 s' A_t^{-1} s\} \exp\{-(1/2) f_t s' A_t^{-1} x_t x_t' A_t^{-1} s\} \\ &= \exp\{is'\hat{\beta}_{t-1} - (1/2)\sigma^2 s' A_{t-1} s\} \\ &= cf_{t-1}(s) . \end{aligned}$$

This shows that under the measure Q_t the characteristic function $cf_t(s)$ satisfies the martingale property. It follows that $cf_t(s)$ and hence the posterior distribution Π, are local Q_t-martingales.

2.7. Remarks

(i) Theorem 2.5(a) tells us that $\hat{\beta}_t$ evolves as a martingale under the measure Q_t. Part (b) of the theorem shows that $\hat{\beta}_t$ is a Gaussian process with conditional variance matrix

$$\mathrm{var}(\hat{\beta}_t|\mathcal{F}_{t-1}) = \sigma^2(A_{t-1}^{-1} - A_t^{-1}) = f_t A_t^{-1} x_t x_t' A_t^{-1} .$$

This is the conditional variance of the martingale difference $\hat{\beta}_t - \hat{\beta}_{t-1}$. Observe that this variance matrix is singular when $k > 1$. In the time period from $t-1$ to t only one additional observation, viz. x_t, on the regressor is available. The increase in precision with which $\hat{\beta}_t$ is determined depends on this extra observation and is measured by $\sigma^2(A_{t-1}^{-1} - A_t^{-1})$ $= f_t A_t^{-1} x_t x_t' A_t^{-1}$. As $t \to \infty$, the variance matrix $\sigma^2(A_{t-1}^{-1} - A_t^{-1}) \to 0$ when the excitation condition $\lambda_{\min}(A_t) \to \infty$ holds. In this case $\hat{\beta}_t \to_{a.s.} \beta$ and the conditional distribution $\beta_t|\mathcal{F}_{t-1}$ converges to a point process with unit mass at β.

(ii) Part (c) of the theorem shows that the posterior distribution Π_t also evolves like a martingale under Q_t. As new information accumulates the mean and variance matrix of the Gaussian measure Π_t evolve according to the processes $(\hat{\beta}_t)$ and $(\sigma^2 A_t^{-1})$. The best estimate (under the Bayes measure Q_t) of the posterior Π_t given $\mathcal{F}_{t-1}$ is simply Π_{t-1}.

3. MODEL SELECTION AND HYPOTHESIS TESTING

Trajectory dependent location models like the Bayes model (15) can be expected to be hard models to beat, precisely because they rely so intimately on the existing data record. One important element in assessing how satisfactory such models are in practical applications is the dimension of the respective parameter spaces that they require. In fitting the historical time series trajectory improvements in fit are (almost) always possible by raising the dimension of the parameter space. This is certainly true in the present context where $y_{t|t-1} = \hat{\beta}_{t-1}' x_t$ in (15) is the best predictor of y_t given data in $\mathcal{F}_{t-1}$.

General model selection principles introduce penalties for increasing the number of estimated parameters. Among the most popular of these in time series contexts are the order estimation criteria AIC of Akaike (1969, 1977) and BIC of Schwarz (1978) and Rissanen (1978). The statistical properties of these criteria have been intensively investigated in both stationary and nonstationary autoregressive and autoregressive-moving average models. Hannan and Deistler (1988, Ch. 5) provide a recent detailed discussion of the subject. A general treatment which is suited to the present context and which establishes strong consis-

tency of order estimates for variants of the BIC criterion applied to the regressor selection problem in stochastic regression models is given in Pötscher (1989). Applied to (1), order estimates by BIC of the dimension k of β are obtained by minimizing the quantity

(C1) $\quad BIC_k = ln(\hat{\sigma}_k^2) + k\ ln(n)/n\ ,$

where $\hat{\sigma}_k^2$ is the maximum likelihood estimate of σ^2 from the model with order k. Consistency of BIC order estimates have been obtained for stationary autoregressive-moving average models by Hannan (1980, 1981) and nonstationary autoregressive and stochastic regressor models by Paulsen (1984), Tsay (1984) and Pötscher (1989).

In our context a natural measure of model adequacy is provided by the data density dQ_n/dP_n given in (12). The principle we use, which seems appealing in this case, is to choose the model with the greatest data density, that is the model with the greatest likelihood ratio or posterior density in the general model class. We now proceed to develop this criterion.

Let Q_n^k be the "Bayes model" measure given by (11) for a model with k parameters (i.e. $\beta \in \mathbb{R}^k$ in (1)). We rewrite (1) and our later notation to incorporate the index "k" signifying the number of regressors. Thus, we set

(18) $\quad Y_n = X_n(k)\beta(k) + E_n\ ,\ \ A_n(k) = X_n(k)'X_n(k)\ ,$

(19) $\quad \hat{\beta}_n(k) = [X_n(k)'X_n(k)]^{-1}X_n(k)'Y_n\ ,\ \ \hat{E}_n(k) = Y_n - X_n(k)\hat{\beta}_n(k)\ ,\ \ SS_k = \hat{E}_n(k)'\hat{E}_n(k)\ .$

Then

(20) $\quad dQ_n^k/dP_n = \left|(1/\sigma^2)A_n(k)\right|^{-1/2}\exp\left\{(1/2\sigma^2)\hat{\beta}_n(k)'A_n(k)\hat{\beta}_n(k)\right\}$

($k = 1, 2, ..., K$) is a sequence of alternative densities for Y_n taken with respect to the reference measure P_n of the null model in which $Y_n = E_n$. We let K be some maximum number of regressors and suppose that if k_0 is the number of regressors in the true model then $k_0 \le K$. Of course, if $k_0 = \infty$ then we will need to allow $K \to \infty$ as $n \to \infty$ to accommodate this possibility (just as in the application of AIC and BIC model selection principles).

We may proceed directly to maximize (20) over k. However, in general it will be useful to employ a more relevant reference measure than P_n. One such measure with nice properties

is Q_n^K since this corresponds with the "least restricted" option in the class. Multiplying the RN derivatives we obtain

$$
\begin{aligned}
dQ_n^k/dQ_n^K &= (dQ_n^k/dP_n)(dP_n/dQ_n^K) \\
&= \left|(1/\sigma^2)A_n(k)\right|^{-1/2}\left|(1/\sigma^2)A_n(K)\right|^{1/2}\exp\left\{(1/2\sigma^2)[\hat{\beta}_n(k)'A_n(k)\hat{\beta}_n(k) - \hat{\beta}_n(K)'A_n(K)\hat{\beta}_n(K)]\right\} .
\end{aligned} \tag{21}
$$

This expression is the likelihood ratio of the measures corresponding to the two "Bayes models":

$$
\begin{aligned}
&\mathrm{H}(Q_n^k):\ y_{n+1} = \hat{\beta}_n(k)'x_{n+1}(k) + v_{n+1}^k ; \\
&\mathrm{H}(Q_n^K):\ y_{n+1} = \hat{\beta}_n(K)'x_{n+1}(K) + v_{n+1}^K .
\end{aligned}
$$

The first model has k regressors $x_n(k)$; the second, which is the reference model, has K regressors $x_n(K)$. The likelihood ratio dQ_n^k/dQ_n^K measures the support in the data, as it is embodied in the data density, for the restricted model against the base model with K regressors. If we assign equal prior odds to the two models we may actually test $\mathrm{H}(Q_n^k)$ against $\mathrm{H}(Q_n^K)$ using the criterion

(C2) *Accept* $\mathrm{H}(Q_n^k)$ *in favor of* $\mathrm{H}(Q_n^K)$ *if* $dQ_n^k/dQ_n^K > 1$.

This criterion, as we discuss below, gives a completely consistent "Bayes model" test in that the probabilities of both types of error tend to zero as $n \to \infty$.

The following partitioned regression notation will be helpful in formulating some alternative representations of (21).

$$
\begin{aligned}
&Y_n = X_n(K)\beta(K) + E_n = X_n(k)\beta(k) + X_n(*)\beta(*) + E_n , \\
&A_n(*) = X_n(*)'X_n(*) , \\
&A_n(*, k) = X_n(*)'X_n(k) , \\
&A_n(**.k) = A_n(*) - A_n(*, k)A_n(k)^{-1}A_n(k, *) , \\
&\hat{\beta}_n(*) = \left[X_n(*)'X_n(*) - X_n(*)'X_n(k)(X_n(k)'X_n(k))^{-1}X_n(k)'X_n(*)\right]^{-1} \\
&\qquad \times \left[X_n(*)'Y_n - X_n(*)'X_n(k)(X_n(k)'X_n(k))^{-1}X_n(k)'Y_n\right] ,
\end{aligned}
$$

$$\bar{\beta}_n(k, K) = \begin{bmatrix} \hat{\beta}_n(k) \\ 0 \end{bmatrix} \begin{matrix} k \\ K-k \end{matrix},$$

$$ss_k = (Y_n - X_n(k)\hat{\beta}_n(k))'(Y_n - X_n(k)\hat{\beta}_n(k)) \ , \ ss_K = (Y_n - X_n(K)\hat{\beta}_n(K))'(Y_n - X_n(K)\hat{\beta}_n(K)) \ .$$

The expressions given in Lemma 3.1 below follow from standard regression manipulations.

3.1. LEMMA

$$\frac{dQ_n^K}{dQ_n^k} = \left|(1/\sigma^2)A_n(**.k)\right|^{-1/2} \exp\left\{(1/2\sigma^2)(ss_k - ss_K)\right\} , \tag{22}$$

$$\frac{dQ_n^K}{dQ_n^k} = \left|(1/\sigma^2)A_n(**.k)\right|^{-1/2} \exp\left\{(1/2\sigma^2)\hat{\beta}_n(*)'A_n(**.k)\hat{\beta}_n(*)\right\} , \tag{23}$$

$$\frac{dQ_n^K}{dQ_n^k} = \left|(1/\sigma^2)A_n(**.k)\right|^{-1/2} \exp\left\{(1/2\sigma^2)(\hat{\beta}_n(K) - \bar{\beta}(k,K))'A_n(K)(\hat{\beta}_n(K) - \bar{\beta}_n(k,K))\right\} . \ \square \tag{24}$$

3.2. REMARKS

(i) One advantage of using the reference measure Q_n^K is that it provides a general model for the estimation of the error variance σ^2, giving

$$\hat{\sigma}_K^2 = (1/(n-K))(Y_n - X_n(K)\hat{\beta}_n(K))'(Y_n - X_n(K)\hat{\beta}_n(K)) = ss_K/(n-K) \ ,$$

which is the least squares estimator of σ^2 in (1) when there are K regressors. This estimate of σ^2 can be used in formulae (21)-(24) and leads to our model selection criterion

$$\text{(C3)} \qquad \text{PIC}_k = (dQ_n^K/dQ_n^k)(\hat{\sigma}_K^2) = \left|(1/\hat{\sigma}_K^2)A_n(**.k)\right|^{-1/2} \exp\left\{(1/2\hat{\sigma}_K^2)\hat{\beta}_n(*)'A_n(**.k)\hat{\beta}_n(*)\right\}$$

and order estimator

$$\hat{k} = argmin_k \ \text{PIC}_k \ . \tag{25}$$

In minimizing PIC_k we are maximizing the reciprocal $1/\text{PIC}_k = dQ_n^k/dQ_n^K$ and thereby choosing the model, $\text{H}(Q_n^k)$, that we favor most over $\text{H}(Q_n^K)$ according to the data density.

The properties of this new order estimator will be studied systematically in another paper. We note here that the procedure is asymptotically equivalent to BIC in stationary autoregressive and stationary autoregressive-moving average models. In models with some autoregressive unit roots, however, the PIC criterion (C3) is asymptotically different from BIC. The PIC criterion attaches a greater penalty than BIC for additional regressors in this case, and involves a penalty that is asymptotically of the form $d(k_s, k_u)ln(n)/n$ where $d(k_s, k_u) = k_s + k_u(k_u + 1)$, k_s is the number of stationary regressors and k_u is the number of autoregressive unit roots in the system. The BIC criterion (C1) may be regarded as a specialization of PIC to the case of stationary regressors. Both criteria produce consistent order estimates but as the simulations in Section 5 show, PIC outperforms BIC in terms of correct model choice even in stationary autoregressive systems.

(ii) An alternative version of the PIC criterion can be obtained directly from the conditional "Bayes model" measure (14). We have the density (conditional on $\mathcal{F}_K$)

$$dQ_n^k/d\nu = \text{pdf}(Y_n|\mathcal{F}_K) = (2\pi)^{-(n-K)/2}\exp\left\{-\Sigma_{K+1}^n\left(v_t^k\right)^2/2f_t^k\right\}/\left(\Pi_{K+1}^n f_t^k\right)^{1/2}$$

The idea is now to compare the "Bayes models" $H(Q_n^k)$ and $H(Q_n^K)$ in terms of their densities over the same subsample, viz. $n > K$. We then have (conditional on $\mathcal{F}_K$)

$$dQ_n^k/dQ_n^K = \left(\Pi_{K+1}^n f_t^K/f_t^k\right)^{1/2}\exp\left\{\Sigma_{K+1}^n\left[\left(v_t^K\right)^2/2f_t^K - \left(v_t^k\right)^2/2f_t^k\right]\right\}$$

where $f_t^k = \sigma^2\{1 + x_t(k)'A_{t-1}(k)^{-1}x_t(k)\}$. If we now estimate σ^2 using the reference model with K parameters (i.e. the most complex model), this leads us to the following alternative form of the PIC criterion.

$$\text{(C3')} \qquad \text{PIC}_k' = dQ_n^k/dQ_n^K(\hat{\sigma}_K^2) = \left(\Pi_{K+1}^n \hat{f}_t^K/\hat{f}_t^k\right)^{1/2}\exp\left\{\Sigma_{K+1}^n\left[\left(v_t^K\right)^2/2\hat{f}_t^K - \left(v_t^k\right)^2/2\hat{f}_t^k\right]\right\}$$

where

$$\hat{f}_t^k = \hat{\sigma}_K^2(1 + x_t(k)'A_{t-1}(k)^{-1}x_t(k)) \ , \quad \hat{f}_t^K = \hat{\sigma}_K^2(1 + x_t(K)'A_{t-1}(K)^{-1}x_t(K)) \ ;$$

$$v_t^k = y_t - \hat{\beta}_{t-1}(k)'x_t(k) \ , \quad v_t^K = y_t - \hat{\beta}_{t-1}(K)'x_t(K) \ .$$

The form of PIC′ has the distinct advantage (over PIC) that it is invariant to linear transformations of the regressors $x_t(k)$ and $x_t(K)$. Since PIC′ is constructed from the measures of the competing Bayes models over the same subsample of data, the criterion is free from initialization differences and (implicitly) prior distributions on the parameters. The criterion is as close to an "objective" Bayes criterion as we can hope to get -- it is dependent only on the data history over the subsample from $t = K+1, ..., n$.

(iii) When this paper was in the final stages of write up we learnt of some related work on model selection by Wei (1992). Wei suggests a criterion (called "FIC") based on the use of the "Fisher information" $|A_k(k)|$ as a penalty rather than a sample parameter count. Wei's criterion is to select the model (i.e. k) that minimizes

$$\mathrm{FIC}_k = n\hat{\sigma}_k^2 + \hat{\sigma}_K^2 \ln|A_n(k)| ,$$

(see equation (5.1.1) of Wei, 1992). Using (22) and (C3) our criterion can be transformed to

$$\begin{aligned} 2ln(\mathrm{PIC}_k) &= (ss_k - ss_K)/\hat{\sigma}_K^2 - ln(|A_n(**.k)/\hat{\sigma}_K^2|) \\ &= \hat{\sigma}_K^{-2}[ss_k + \hat{\sigma}_K^2 \, ln\,|A_n(k)/\hat{\sigma}_K^2|] - ss_K/\hat{\sigma}_K^2 - ln\,|A_n(K)/\hat{\sigma}_K^2| . \end{aligned}$$

The quantity in square parentheses above is asymptotically equivalent to FIC_k when the penalty term $ln\,|A_n(k)|$ in FIC_k is replaced by the scale invariant term $ln\,|A_n(k)/\hat{\sigma}_k^2|$ (as suggested in Remark 5.3 of Wei (1991)). Wei's justification for the FIC_k criterion is that it is more meaningful to use statistical information that is relevant to the model (i.e. $ln\,|A_n(k)|$) as a penalty rather than the dimension of the parameter space (k). Our justification for PIC_k is that it is actually the posterior odds in favor of the model with k parameters over the reference model in the given class. Our justification shows that PIC (and the asymptotically equivalent FIC) are Bayesian criteria that are founded on the principle that one should choose the mdoel that is most favored by the data *a posteriori* (i.e. in terms of its posterior odds). Development of the respective "Bayes model" measures Q_n^k and Q_n^K for competing models in the class is the essential element in deriving the criterion PIC and this idea is capable of substantial generalization beyond the present context.

(iv) The exponent in (C3) is one half times the quantity

$$W_n = \hat{\beta}_n(*)'A_n(**.k)\hat{\beta}_n(*)/\hat{\sigma}_K^2 ,$$

which is the Wald statistic for testing the hypothesis that $\beta(*) = 0$ in the partitioned model

$$Y_n = X_n(k)\beta(k) + X_n(*)\beta(*) + E_n .$$

Let us now assume that the model is autoregressive or autoregressive with trend as in (2), (3) or (4) with possibly some unit roots. Then, using the asymptotic theory developed in the Park-Phillips (1988, 1989) papers it is easy to show that $W_n = O_p(1)$ as $n \to \infty$ when the null hypothesis

$$H_0(*) : \beta(*) = 0$$

is correct. Now $\hat{\sigma}_K^2 \to_p \sigma^2$ and $|A_n(**.k)| = O_p(n^m)$ for some integer $m \geq K-k$, again under the null. Hence

$$(dQ_n^K/dQ_n^k)(\hat{\sigma}_K^2) \to_p 0$$

as $n \to \infty$ under $H(*)$. On the other hand, when $H(*)$ is false we have $W_n = O_p(n^\ell)$ for some $\ell \geq 1$ and hence $(dQ_n^K/dQ_n^k)(\hat{\sigma}_K^2)$ diverges as $n \to \infty$. Stating this result formally we have:

3.3. THEOREM: *Let the true model* (1) *be autoregressive or autoregressive with trend as in* (2)-(4) *with stable roots and possibly one or more unit roots. A "Bayes model" test of*

$$H_0(*) : \beta(*) = 0, \textit{ against } H_1(*) : \beta(*) \neq 0$$

is based on the criterion

(C4) *Accept* $H_0(*)$ *in favor of* $H_1(*)$ *if* $(dQ_n^K/dQ_n^k)(\hat{\sigma}_K^2) < 1$.

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

3.4. REMARK. The above theory can be applied directly to test for the presence of a unit root. For instance, in model (4) we set $K = p+2$, $k = p+1$, $\beta(*) = h$ (the coefficient of y_{t-1}) and $x_t(k)' = (1, t, \Delta y_{t-1}, ..., \Delta y_{t-p+1})$. The criterion (C4) then determines which of the following two "Bayes models" the data favors:

$$H(Q_n^{K=p+2}) : \Delta y_{n+1} = \hat{h}_n y_n + \Sigma_{i=1}^{p-1} \hat{\varphi}_{in} \Delta y_{n+1-i} + \hat{\mu}_n + \hat{\gamma}_n(n+1) + \varepsilon_{n+1}$$

or

$$H(Q_n^{k=p+1}) : \Delta y_{n+1} = \Sigma_{i=1}^{p-1} \tilde{\varphi}_{in} \Delta y_{n+1-i} + \tilde{\mu}_n + \tilde{\gamma}_n(n+1) + \varepsilon_{n+1} .$$

The second model explicitly incorporates a unit root. Both have data-determined and time-evolving coefficients. The only reason for not preferring $H(Q_n^{K=p+2})$ is that it carries the cost of an additional parameter. The criterion (C4) assesses this cost through the penalty that is incorporated in the denominator of $(dQ_n^K/dQ_n^k)(\hat{o}_K^2)$. If this cost outweighs the gain that is measured in terms of the reduction in the error sum of squares (i.e. $(ss_k - ss_K)/\hat{o}_K^2$, as in the formulation (22)) from the inclusion of the additional regressor then the criterion favors the presence of a unit root in the model. When the cost does not outweigh the gain, then $H(Q_n^{K=p+2})$ is chosen and the unit root rejected.

4. TESTING DIFFERENCE STATIONARITY VERSUS TREND STATIONARITY

Bayes model tests of the form given in (C4) may be used to test any sharp null hypothesis (like that of a unit root) about a model. As our discussion makes clear, the test compares one Bayes model with another and the outcome of the test depends on the balance of the gains versus the costs of additional regressors. The testing apparatus will be employed in this section to assess evidence for trend versus difference stationarity. In such cases we must give attention to the gains and costs of including a deterministic trend as well as additional lagged variables. As pointed out in the Introduction, our approach as well as our methodology is different in this respect from existing work on this problem.

Our Bayesian model selection criteria will be used to determine not only the stochastic regressor components of the model but also the form of any deterministic polynomial trend. We therefore do not maintain the presence of a polynomial trend function in the model as in (4) but instead rely on our selection criteria to decide whether it should be included and, if so, in what form.

In addition, we wish to allow for more general stochastic regressor models than

autoregressions. The reference model we use for assessing evidence in favor of the presence of a unit root is the ARMA(p, q) + trend model given earlier in (5) (and repeated here for convenience):

$$(5) \qquad \Delta y_t = h y_{t-1} + \Sigma_{i=1}^{p-1} \varphi_i \Delta y_{t-i} + \Sigma_{j=1}^{q} \psi_j \varepsilon_{t-j} + \mu + \gamma t + \varepsilon_t ,$$

where ε_t = iid $N(0, \sigma^2)$. To deal with the MA error components in (5) we propose employing the first two stages of the so-called Hannan-Rissanen (1982, 1983) procedure. This involves the use of a (possibly) long first stage autoregression in place of (5) to estimate the error process ε_t. The residuals from the first stage autoregression are then used to replace the lagged errors ε_{t-j} in the MA component of (5) in the second stage. In both first and second stages a consistent model selection procedure (like BIC or PIC) is used to determine the appropriate order of the regression. In addition, we apply a model selection procedure to determine the order of the accompanying trend polynomial.

The precise steps in our procedure are detailed in full in the algorithm that follows:

4.1. Model Selection Algorithm and Data-Based Unit Root Test

Step 1. Set maximum orders for the stochastic and deterministic components of the model as follows:

K = maximum autoregressive lag,

J = maximum moving average lag,

L = maximum degree of polynomial time trend.

Step 2. Run a sequence of long autoregressions with a fitted time trend of the form

$$(26) \qquad \Delta y_t = \hat{a}_0 y_{t-1} + \Sigma_{i=1}^{k-1} \hat{a}_i \Delta y_{t-i} + \Sigma_{t=0}^{L} \hat{b}_t t^t + \text{residual}$$

for $k = 0, 1, \ldots, \bar{K}$. Choose $\bar{K} > K$ to be large in this step (for example, we might choose $\bar{K} = 10$ or 15 when the sample size $n = 100$). (When $k = 0$ the regression is "$y_t = \Sigma_{t=0}^{L} \hat{b}_t t^t$ + residual" and there is no autoregressive component.) Select the order of the autoregression using the PIC or BIC criteria. If PIC is used it is helpful to run regression (26) with $k = \bar{K}$ first to set the reference measure. Let $\hat{k}$ be the selected order of this regression.

STEP 3. Run a sequence of autoregressions with fitted time trends of variable degree of the form

$$(27) \qquad \Delta y_t = \hat{a}_0 y_{t-1} + \Sigma_{i=1}^{k-1} \hat{a}_i \Delta y_{t-i} + \Sigma_{j=0}^{\ell} \hat{b}_j t^j + \text{residual}$$

for $\ell = -1, 0, 1, ..., L$. (When $\ell = -1$, no intercept is included in (27).) Select the order of the trend polynomial using the PIC or BIC criterion. (Again, if PIC is used, run the regression (27) with $\ell = L$ first to set the reference measure.) Let $\hat{\ell}$ be the selected degree of the polynomial time trend from this regression. Compute the residuals from this regression and call them $\hat{\varepsilon}_t$.

STEP 4. Run a double sequence (or array) of regressions with lagged residual regressors $\hat{\varepsilon}_{t-s}$ of the form

$$(28) \qquad \Delta y_t = \hat{a}_0 y_{t-1} + \Sigma_{i=1}^{k-1} \hat{a}_i \Delta y_{t-i} + \Sigma_{j=0}^{\hat{\ell}} \hat{b}_j t^j + \Sigma_{s=1}^{q} \hat{c}_s \hat{\varepsilon}_{t-s} + \text{residual}$$

for ($k = 0, 1, ..., \bar{k} = \max(\hat{k}, K)$; $q = 0, 1, ..., J$). Select the orders of the moving average and autoregressive components simultaneously from this array of regressions using the PIC or BIC criterion. If PIC is used, the regression (28) can be run with $k = \bar{k}$ and $q = J$ first to set the reference measure. If BIC is used it is preferable to compute the estimate of σ^2 that is utilized in the criterion (C1) from the second stage regression (28) by means of the recursion

$$\varepsilon_t^* = \Delta y_t - \hat{a}_0 y_{t-1} - \Sigma_{i=1}^{\bar{k}} \hat{a}_i \Delta y_{t-i} - \Sigma_{j=0}^{\hat{\ell}} \hat{b}_j t^j + \Sigma_{s=1}^{q} \hat{c}_s \varepsilon_{t-s}^*$$

with the initialization $\varepsilon_t^* = y_t = 0$ for $t \le 0$. Kavalieris (1991) shows that this method leads to improved estimates of the autoregressive-moving average orders. Let $\hat{p}$ and $\hat{q}$ be the respective order of the autoregressive and moving average components selected from this array of regressions.

STEP 5. Run the regression (28) with selected orders $\hat{p}$, $\hat{q}$, $\hat{\ell}$ of the autoregressive, moving average and trend components giving

$$(29) \qquad \Delta y_t = \bar{a}_0 y_{t-1} + \Sigma_{i=1}^{\hat{p}-1} \bar{a}_i \Delta y_{t-i} + \Sigma_{j=0}^{\hat{\ell}} \bar{b}_j t^j + \Sigma_{s=1}^{\hat{q}} \bar{c}_s \hat{\varepsilon}_{t-s} + \bar{\varepsilon}_t .$$

STEP 6. (a) If $\hat{p} > 0$, compute the "Bayes model" test criterion (C4) to assess the support for the presence of a unit root in the selected model (29). Writing (29) in the regression notation

$$\Delta y = \alpha_0 y_{-1} + Z\delta + \varepsilon ,$$

the test criterion has the simple form

$$(30) \qquad BLR(\alpha_0) = (dQ_n^M/dQ_n^{M-1})(\hat{\sigma}_K^2) = \left\{(1/\hat{\sigma}_K^2) y_{-1}' \bar{P}_Z y_{-1}\right\}^{-1/2} \exp\left\{(1/2\hat{\sigma}_K^2)\alpha_0^2 (y_{-1}' \bar{P}_Z y_{-1})\right\} ,$$

where $M = \hat{p} + \hat{q} + \hat{\ell} + 1$, $\bar{P}_Z = I - Z(Z'Z)^{-1}Z'$, and $\hat{\sigma}_K^2 = \varepsilon'\varepsilon/(n-M)$. The Bayes model with a unit root is favored over (29) when $BLR(\alpha_0) < 1$.

(b) If $\hat{p} = 0$, we accept that there is no autoregressive component in the model and, hence, no autoregressive unit root. □

The order estimators $(\hat{p}, \hat{q}, \hat{\ell})$ are consistent estimators of the true orders, provided the latter are finite (see Hannan (1980), Hannan-Rissanen (1982), Hannan-Diestler (1988, Ch. 5) and Kavalieris (1991) for details in the case of the BIC criterion). Also the fitted residuals $\hat{\varepsilon}_{t-s}$ are consistent estimators of ε_{t-s} in (29). (If the true model has an MA component the residuals are still consistently estimated in the first stage regression (26) with $k = K$ because K is selected to be large and allowed to tend to infinity with n, so that the selected AR order $\hat{k} \to \infty$ in this case.) It follows that (29) is asymptotically correctly specified and $(\alpha_0 - a_0)^2 (y_{-1}' \bar{P}_Z y_{-1}) = O_p(1)$ where a_0 is the true value of α_0. Furthermore, $y_{-1}' \bar{P}_Z y_{-1}$ diverges as $n \to \infty$ so that the statistic $\mathrm{BLR}(a_0) \to_p 0$ as $n \to \infty$ when $a_0 = 0$, whereas $\mathrm{BLR}(a_0)$ diverges to infinity when $a_0 \neq 0$. We have:

4.2. THEOREM. *The Bayes model likelihood ratio criterion* $\mathrm{BLR}(a_0)$ *given by* (30) *provides a completely consistent test of the hypothesis of a unit root in the class of finite parameter* $\mathrm{ARMAX}(p, q, \ell)$ *models with trend polynomials as exogenous regressors.*

4.3. REMARKS. (i) The sequence of model selecting regressions in Algorithm 4.1 can be used for a variety of ultimate purposes. Here we have focussed on obtaining a data-based unit root test using our "Bayes model" test criterion (C4). This relies on the $\mathrm{BLR}(a_0)$ statistic given in

(30). We could also use the algorithm to extract a Dickey-Fuller type t-test for a unit root as Said-Dickey (1984) do in terms of a long autoregression (leading to the so-called ADF or augmented Dickey-Fuller test); but in our procedure a general ARMAX system would be considered and our order selection methods would be employed to determine the best model before attempting to test for the presence of a unit root. Work on this alternative procedure is now proceeding and will be explored in a later paper.

(ii) A major difference between our procedure and the ADF procedure as it is used in practice is that we allow the data to select the order of the trend polynomial. As a result, our test procedure is not invariant to the presence of trend or the trend coefficients themselves. We view this as an advantage. Indeed, the stronger is the trend the more likely is our procedure to find a trend to be present in the data in finite samples. Moreover, if the true model has a unit root with drift then our procedure will asymptotically select a model with a unit root and intercept i.e. model (29) with $a_0 = 0$ and $\hat{\ell} = 0$. In this sense, our statistical procedure will be economical on the number of parameters and, unlike the Dickey-Fuller procedure, will not always include parameters and variables that are unnecessary under the null hypothesis.

(iii) We may also be interested in the "Bayes model" produced by 4.1, for example, if we wish to use the model for forecasting. In such cases we may wish to proceed with the third stage of the Hannan-Rissanen procedure (see Hannan-Rissanen (1982) and Hannan-Deistler (1988, Ch. 5)) in order to obtain asymptotically efficient estimates of the coefficients. □

All of the procedures outlined in this section of the paper have been programmed in GAUSS-386i (Version 2.2). When $n = 100$, $\bar{K} = 10$, $K = 3$, $J = 3$, $L = 1$ the computation time taken by the algorithm on a 486-33 PC is approximately 0.85 seconds.

5. SIMULATION EVIDENCE

Simulations were conducted to evaluate the performance of our model selection criterion PIC and our data-based algorithm for detecting the presence of a unit root. We shall discuss these experiments in turn.

(a) *Model selection by PIC*

The model chosen for this experiment was the AR(p) given by (2) with $p = 3$ and $\varepsilon_t \sim$ iid $N(0, 1)$. In the PIC criterion the reference model used in the construction of PIC_k in (C3) was an AR(K) with $K = 10$. The BIC criterion given in (C1) and the AIC criterion of Akaike (1969) were also used for comparative purposes. A sample size of $n = 100$ was used and a large grid of autoregressive coefficients were considered giving a range of models from three unit roots to a nearly iid process. In terms of the roots (λ_i, $i = 1, 2, 3$) of the characteristic equation of the autoregression we took a grid of values in increments of 0.20 for each root λ_i in the interval $[-0.8, 1]$.

- - - - - - - - - - - - - - -
Figure 1 about here
- - - - - - - - - - - - - - -

The results are graphed in Figure 1, which gives a surface that displays the difference between the estimated probability (based on 10,000 replications) of a correct model choice by PIC and BIC, respectively. The surface shows that PIC outperforms BIC almost uniformly over the parameter space considered here. The probability of a correct model choice by PIC exceeds that of BIC by over 0.06 in some cases and on average by more than 0.02. Observe that PIC improves BIC not only in the nonstationary and near-nonstationary region but also in the stationary region of the parameter space.

Table 1 gives detailed model choice statistics for AIC, BIC and PIC for eight specific parameter configurations of $(\lambda_1, \lambda_2, \lambda_3)$ including stationary and nonstationary cases. Only in the case where $(\lambda_1, \lambda_2, \lambda_3) = (0.4, 0.4, 0.4)$ does BIC choose the correct model more times than PIC and then the estimated difference P(correct model choice by BIC) - P(correct model choice by PIC) = 0.005. In this case all the criteria favor the more parsimonious AR(2) over

TABLE 1: MODEL SELECTION BY AIC, BIC AND PIC IN AN AR(3)

Lag order chosen in AR(p)	$\lambda_1 = \lambda_2 = \lambda_3 = 1.00$			$\lambda_1 = \lambda_2 = \lambda_3 = 0.80$			$\lambda_1 = \lambda_2 = \lambda_3 = 0.60$			$\lambda_1 = \lambda_2 = \lambda_3 = 0.40$		
	AIC	BIC	PIC	AIC	BIC	PIC	AIC	BIC	PIC	AIC	BIC	PIC
1	0	0	0	0	0	0	0	0	0	108	453	381
2	0	0	0	17	63	60	3462	6178	5775	7476	8924	9063
3	7904	9537	9630	8014	9492	9626	4974	3557	4014	1449	537	483
4	1123	387	303	1123	365	261	841	223	180	453	64	51
5	464	57	54	394	59	34	337	33	21	225	13	18
6	218	13	8	229	16	16	186	7	9	123	7	4
7	117	6	3	105	5	3	87	2	0	75	1	0
8	82	0	0	55	0	0	58	0	1	46	1	0
9	60	0	2	41	0	0	39	0	0	29	0	0
10	32	0	0	22	0	0	16	0	0	16	0	0
	$\lambda_1 = 1.00$, $\lambda_2 = 0.60$, $\lambda_3 = 0.40$			$\lambda_1 = 1.00$, $\lambda_2 = 0.60$, $\lambda_3 = 0.20$			$\lambda_1 = 1.00$, $\lambda_2 = 0.60$, $\lambda_3 = 0.00$			$\lambda_1 = 0.80$, $\lambda_2 = 0.80$, $\lambda_3 = 0.40$		
	AIC	BIC	PIC	AIC	BIC	PIC	AIC	BIC	PIC	AIC	BIC	PIC
1	0	0	0	0	0	0	0	0	0	0	0	0
2	2650	5163	4770	6226	8550	8449	8004	9582	9637	2332	4717	4319
3	5621	4543	4992	2551	1319	1429	1101	340	299	5916	4984	5427
4	940	242	186	610	104	93	431	54	46	968	255	205
5	377	41	40	270	18	22	197	19	11	376	36	37
6	172	8	10	161	9	7	105	2	2	172	6	10
7	107	2	0	85	0	0	69	0	0	107	2	1
8	60	1	2	46	0	0	44	0	0	56	0	1
9	44	0	0	30	0	0	28	0	0	41	0	0
10	29	0	0	21	0	0	21	0	0	32	0	0

Notes: Number of replications = 10,000; sample size n = 100

an AR(3) and BIC underestimates the order by choosing an AR(1) more frequently than PIC. For the other parameter values PIC clearly dominates BIC in terms of correct model choice by as much as 10% in some instances (e.g. $\lambda_1 = 0.80$, $\lambda_2 = 0.80$, $\lambda_3 = 0.40$). The tendency towards overestimation of model order by AIC is evident in all cases.

(b) *Posterior odds data-based tests for a unit root*

Tables 2(a) and 2(b) show estimated rejection probabilities for the data-based unit root test described in the algorithm of Section 4.1. An ARMA($\hat{p}$, $\hat{q}$) + trend ($\hat{\ell}$) model is constructed with estimated orders $\hat{p}$, $\hat{q}$ and $\hat{\ell}$ using Steps 1-6 of the algorithm. The estimated rejection probabilities given in these tables are obtained from 1,000 replications. The unit root test is based on the posterior odds ratio given by (30) when $\hat{p} > 0$. The odds favor the presence of a unit root when the statistic satisfies BLR(π_0) < 1. When $\hat{p} = 0$, a unit root model is automatically rejected as there is no autoregressive component in the selected model.

TABLE 2(a): SIMULATION ESTIMATES OF UNIT ROOT REJECTION PROBABILITIES

Sample Size	ARMA(1,1) model: $y_t = ay_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$						AR(2) model: $y_t = ay_{t-1} + u_t$, $u_t = \rho u_{t-1} + \varepsilon_t$					
	θ	1.00	0.95	0.90	0.85	0.80	ρ	1.00	0.95	0.90	0.85	0.80
	-0.80	0.173	0.664	0.954	0.998	1.000	-0.80	0.021	0.179	0.605	0.884	0.982
	-0.60	0.058	0.291	0.561	0.786	0.911	-0.60	0.017	0.170	0.531	0.838	0.955
	-0.40	0.048	0.271	0.536	0.702	0.813	-0.40	0.026	0.157	0.456	0.738	0.851
	-0.20	0.040	0.268	0.641	0.846	0.923	-0.20	0.036	0.230	0.612	0.843	0.900
$n=100$	0.00	0.020	0.118	0.402	0.774	0.945	0.00	0.020	0.112	0.437	0.798	0.960
	0.20	0.019	0.112	0.336	0.596	0.828	0.20	0.009	0.095	0.274	0.582	0.788
	0.40	0.042	0.233	0.579	0.805	0.902	0.40	0.017	0.091	0.310	0.573	0.766
	0.60	0.076	0.340	0.720	0.886	0.932	0.60	0.011	0.059	0.193	0.375	0.630
	0.80	0.095	0.435	0.767	0.887	0.920	0.80	0.009	0.024	0.074	0.170	0.298
	-0.80	0.064	0.455	0.883	0.995	1.000	-0.80	0.015	0.267	0.817	0.989	1.000
	-0.60	0.022	0.170	0.377	0.589	0.826	-0.60	0.012	0.255	0.782	0.981	1.000
	-0.40	0.020	0.220	0.424	0.576	0.681	-0.40	0.015	0.214	0.737	0.978	1.000
	-0.20	0.028	0.346	0.750	0.865	0.898	-0.20	0.020	0.323	0.816	0.962	0.999
$n=150$	0.00	0.010	0.175	0.691	0.972	0.995	0.00	0.017	0.181	0.703	0.971	0.999
	0.20	0.018	0.186	0.575	0.866	0.977	0.20	0.011	0.095	0.460	0.807	0.976
	0.40	0.035	0.408	0.812	0.954	0.982	0.40	0.006	0.119	0.466	0.802	0.958
	0.60	0.058	0.576	0.906	0.940	0.969	0.60	0.001	0.085	0.351	0.683	0.864
	0.80	0.085	0.649	0.882	0.949	0.944	0.80	0.003	0.031	0.154	0.352	0.468

Number of replications = 1,000

TABLE 2(b): SIMULATION ESTIMATES OF UNIT ROOT REJECTION PROBABILITIES

Sample Size	ARMA(1,1) + trend model: $y_t = dt + x_t$, $d = 0.025$, $x_t = ay_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$						AR(2) + trend model: $y_t = dt + x_t$, $d = 0.025$, $x_t = ay_{t-1} + u_t$, $u_t = \rho u_{t-1} + \varepsilon_t$					
	θ	1.00	0.95	0.90	0.85	0.80	ρ	1.00	0.95	0.90	0.85	0.80
	-0.80	0.804	0.906	0.987	0.999	1.000	-0.80	0.188	0.216	0.615	0.906	0.982
	-0.60	0.371	0.486	0.700	0.874	0.949	-0.60	0.164	0.248	0.567	0.851	0.968
	-0.40	0.325	0.409	0.544	0.688	0.862	-0.40	0.184	0.225	0.512	0.756	0.861
	-0.20	0.289	0.370	0.542	0.654	0.817	-0.20	0.246	0.347	0.672	0.866	0.936
$n=100$	0.00	0.147	0.167	0.323	0.477	0.652	0.00	0.177	0.208	0.458	0.803	0.961
	0.20	0.162	0.168	0.322	0.420	0.528	0.20	0.142	0.137	0.356	0.562	0.803
	0.40	0.226	0.303	0.529	0.668	0.781	0.40	0.154	0.154	0.314	0.632	0.796
	0.60	0.279	0.439	0.643	0.831	0.855	0.60	0.129	0.143	0.241	0.442	0.647
	0.80	0.310	0.470	0.726	0.832	0.857	0.80	0.100	0.101	0.149	0.212	0.299
	-0.80	0.489	0.769	0.972	0.998	1.000	-0.80	0.131	0.192	0.439	0.728	0.931
	-0.60	0.147	0.285	0.514	0.755	0.902	-0.60	0.122	0.219	0.398	0.686	0.903
	-0.40	0.148	0.261	0.462	0.631	0.793	-0.40	0.131	0.187	0.325	0.594	0.786
	-0.20	0.213	0.354	0.594	0.771	0.884	-0.20	0.175	0.312	0.514	0.696	0.861
$n=150$	0.00	0.090	0.205	0.340	0.609	0.845	0.00	0.102	0.192	0.330	0.582	0.815
	0.20	0.123	0.195	0.328	0.508	0.653	0.20	0.103	0.184	0.266	0.443	0.592
	0.40	0.202	0.357	0.619	0.780	0.856	0.40	0.107	0.169	0.364	0.581	0.703
	0.60	0.227	0.476	0.750	0.841	0.850	0.60	0.098	0.117	0.329	0.498	0.652
	0.80	0.252	0.599	0.814	0.885	0.843	0.80	0.071	0.069	0.198	0.324	0.462

Number of replications = 1,000

For the first case, given in Table 2(a), we considered models with no trend and the maximum degree of the polynomial trend was set at $L = -1$ in Step 1 of the algorithm. Results for both ARMA(1,1) and AR(2) models are shown in the table. For the ARMA (1,1) model and $n = 100$ the rejection probability under the null (when $a = 1.00$) is in the range 0.06-0.09 for all values of the MA coefficient θ except $\theta = -0.8$. At $\theta = -0.8$ the rejection probability is 0.173 and in this case the algorithm more frequently selects a model with no autoregressive (and moving average) component giving $\hat{p} = 0$. The rejection probabilities rise rapidly for all values of θ as a departs from unity. When $n = 150$, the rejection probabilities at the null are noticeably smaller than when $n = 100$, corresponding to the fact that the type I error for this test goes to zero as n increases. Again, rejection probabilities (or power) increase rapidly for $a \neq 1$. Similar observations apply to the AR(2) model. In this case the size characteristics of the test are more stable at $n = 100$, although rejection under the null decreases as the second AR coefficient $\rho \rightarrow 1.00$, as would be expected since there are now almost two unit roots in the model. The power in this case is substantial for all parameter values except $\rho = 0.80$ (again, since the second root is near to unity). Size decreases and power increases when n increases to 150, as in the ARMA(1,1) case.

Table 2(b) shows corresponding results for the same models and parameter values but has a linear trend in the generating mechanism and allows the algorithm to select the trend degree as well as the lag orders. In this case, the rejection probabilities under the null are much higher than in Table 2(a), as would be expected. In finding the best model for the data, the algorithm frequently favors a Bayes model with evolving trend and transient dynamic coefficients even when there is a unit root. When n increase from 100 to 150 the probability of rejecting the presence of a unit root under the null falls and the procedure clearly has more discriminating power for samples of this size. For the AR(2) + trend model the rejection probabilities under the null are lower and more uniform than for the ARMA(1,1) + trend model. Power increases rapidly in this case except when $\rho = 0.80$, in which case the presence of the two positive and large autoregressive roots leads to somewhat lower rejection probabilities (e.g. 0.46 for $a = \rho = 0.8$ when $n = 150$).

Overall, the authors find the results of these simulation exercises to be quite encouraging. Even for cases where classical unit root tests have serious size distortion, like the ARMA(1,1) model with a strong negative moving average effect, the data-based procedure seems to work quite well. As in the case of classical tests, the presence of trends generally reduces discriminatory power but when $n = 150$ the data-based procedure gives results that we find very acceptable, especially given the complexity of this problem and the disappointing performance of other methods. For AR models with trend the procedure seems to work rather well.

6. EMPIRICAL ILLUSTRATION

The methods of Sections 3 and 4 were applied to the fourteen historical US time series studied by Nelson-Plosser (1982). For each of the fourteen series we applied the algorithm described in Section 4. We set the maximum polynomial time trend degree at $L = 1$ in Step 1, the long autoregressive lag parameter in Step 2 at $\bar{K} = 10$ and the maximum ARMA lag lengths in Step 4 at $K = J = 3$. Both ARMA + trend and AR + trend models were employed. We used our model selection criterion PIC to choose the trend degree and AR order and the BIC criterion to choose the ARMA lag lengths in Step 4. The BIC criterion was used in the latter step as it was simpler to program. In later work we plan to use the PIC criterion throughout, once the software is written to accommodate ARMA specifications.

The empirical results are shown in Table 3. Real GNP, nominal GNP and per capita GNP are all found to have AR(2) specifications with no deterministic trend but with a unit autoregressive root. The posterior odds in favor of the unit root for these series ranges from 60:1 to 400:1. Note that these are odds in favor of the "Bayes model" with a unit root against the corresponding "Bayes model" without the unit root -- see Remark 3.4 above. This result provides strong Bayesian confirmation of the earlier classical results of Nelson-Plosser.

Only two series (money stock and industrial production) are found to have linear deterministic trends. For these two series the odds against the presence of a unit root in the "Bayes models" are around 6:1 and 3:1 respectively. Unemployment is found to be stationary with a nonzero mean (i.e. the model with fitted intercept is selected in favor of a model with a

linear trend and a model with no trend and no intercept).

TABLE 3: EMPIRICAL RESULTS FOR NELSON-PLOSSER DATA

Series	Block A Model class = ARMA(p, q) + linear trend				Block B Model class = AR(p) + linear trend			
	Model selected		Long-run autoregressive coefficient	Posterior odds in favor of a unit root = 1/BLR(a_0)	Model Selected		Long-run autoregressive coefficient	Posterior odds in favor of a unit root = 1/BLR(a_0)
	Dynamics	Deterministic trend degree*			Dynamics	Deterministic trend degree*		
Real GNP	AR(2)	-1	1.003	59.523	AR(2)	-1	1.003	59.523
Nominal GNP	AR(2)	-1	1.002	64.516	AR(2)	-1	1.002	64.516
Real p.c. GNP	AR(2)	-1	1.001	400.000	AR(2)	-1	1.001	400.000
Industrial production	AR(1)	1	0.841	0.169	AR(1)	1	0.841	0.169
Employment	AR(2)	-1	1.001	129.870	AR(2)	-1	1.001	129.870
Unemployment	ARMA(1,1)	0	0.585	0.000	AR(4)	0	0.709	0.019
GNP deflator	AR(2)	-1	1.003	69.444	AR(2)	-1	1.003	69.444
Consumer prices	ARMA(2,1)	-1	1.002	172.414	AR(6)	-1	1.001	555.555
Nominal wages	ARMA(1,1)	-1	1.005	0.001	AR(2)	-1	1.002	40.186
Real wages	AR(1)	-1	1.004	11.001	AR(2)	-1	1.004	11.001
Money Stock	AR(2)	1	0.916	0.321	AR(2)	1	0.916	0.321
Velocity	AR(1)	-1	0.981	4.472	AR(1)	-1	0.981	4.472
Bond yields	AR(1)	-1	1.019	12.642	AR(1)	-1	1.019	12.642
Stock prices	AR(1)	-1	1.007	81.301	AR(1)	-1	1.007	81.301

*Legend: deterministic trend degree p = -1 (no trend or intercept), p = 0 (intercept only), p = 1 (intercept + linear trend)

All of the remaining series are found to be stochastically nonstationary. The nominal wage series provides a very interesting case where a "Bayes model" with a mildly explosive long run autoregressive coefficient of 1.0054 is selected over a model with a unit root. The nonstationary models selected for all of the other series have unit roots. Note that in the case of the stock price series the odds in favor of the presence of a unit root are close to 100:1.

Table 3 also details the models and the lag orders selected. We note that eleven of the series are found to be autoregressive, either AR(2) (6 series) or AR(1) (5 series). Three of the series are autoregressive moving average, either ARMA(1,1) (unemployment and the nominal wage) or ARMA(2,1) (the consumer prices).

Results obtained by restricting the model class to be purely autoregressive are given in Block B of Table 3. All of the conclusions concerning stochastic nonstationarity are the same. The only important change from restricting the class of models to be autoregressive is for the nominal wage series. In this case an AR(1) model with no deterministic trend is selected (as distinct from an ARMA(1,1)) and the posterior odds are in favor of the presence of a unit root as compared with a mildly explosive autoregressive root when the wider model class is used for model selection.

7. CONCLUSION

This paper puts forward what we believe is a new paradigm for Bayesian inference in time series. As we have shown in Theorem 2.3, the effect of data conditioning in a time series model is to alter the context of statistical inference from the original model to a location model, like (15), where $\hat{\beta}_n' x_{n+1}$ is the best estimate of the location of y_{n+1} given the historical trajectory (i.e. information in $\mathcal{F}_n$). In such a context, a Gaussian (or asymptotically Gaussian) posterior density for the parameter vector β that is centered on the maximum likelihood estimate $\hat{\beta}_n$ seems eminently reasonable, whereas it is much less reasonable in the context of the original time series model because of the poor sampling properties of $\hat{\beta}_n$. We call location models such as (15) "Bayes models" because they arise naturally in the passage to the posterior density due to data conditioning. Associated with these models are probability measures that we call "Bayes model" measures. Our new paradigm for Bayesian inference works explicitly with these "Bayes model" measures. The measures allow us to compare models and to conduct tests, almost as we do in classical theory, by using likelihood ratios. What is especially powerful in the new paradigm is that problems that have been separately treated before in time series analysis (like model selection and hypothesis testing) are now treated simply as different aspects of the same common theory. Thus, a model is selected when its density under these new measures is the largest in a given class. And a point null hypothesis is favored by the data when the likelihood ratio of the "Bayes model" densities exceeds given prior odds, which we typically set to unity. In other words, the likelihood ratio or RN derivative of the respective "Bayes model" measures allows us to discriminate statistically equally well among models and among sharp hypotheses about parameters in those models.

The empirical results of Section 6 provide support for the earlier conclusions of Nelson-Plosser (1982) concerning the presence of stochastic trends in US historical time series. Especially interesting in these results is the fact that deterministic trends receive support from model selection methods for only two series (the money stock and industrial production). These series and the unemployment rate are the only series found to be trend or level stationary.

We emphasize that our approach to inference is very different from both classical and Bayesian methods that have heretofore been adopted in studies relating to the presence or absence of unit roots in economic time series. Since our methods are data-based and integrally involve model selection we allow the data to choose the most appropriate model. As more data accumulates, this approach recognizes the potential need for the model itself to evolve. And when the model changes, so too may the conclusions concerning the presence or absence of stochastic nonstationarity. We view this flexibility and updating as an inherent advantage of our approach.

This approach to Bayesian inference in time series models has many applications beyond those presented here. The authors plan to report on analytical extensions of the theory to nonlinear models, multivariate models and models with cointegrated processes in later work.

8. REFERENCES

Akaike, H. (1969). "Fitting autoregressive models for prediction," *Annals of the Institute of Statistical Mathematics* 21, 243-247.

Akaike, H. (1977). "On entropy maximization principle," in P. R. Krishnarah (ed.), *Applications of Statistics*. Amsterdam: North-Holland, pp. 27-41.

Durbin, J. (1960). "The fitting of time series models," *International Statistical Review*, 28, 233-244.

Hannan, E. J. (1980). "The estimation of the order of an ARMA process," *Annals of Statistics*, 8, 1071-1081.

Hannan, E. J. (1981). "Estimating the dimension of a linear system," *Journal of Multivariate Analysis*, 11, 459-473.

Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. New York: John Wiley & Sons.

Hannan, E. J. and J. Rissanen (1982). "Recursive estimation of ARMA order," *Biometrika*, 69, 273-280 [Corrigenda, *Biometrika*, 1983, 70].

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge, UK: Cambridge University Press.

Kavalieris, L. (1991). "A note on estimating autoregressive-moving average order," *Biometrika*, 78, 920-922.

Nelson, C. R. and C. Plosser (1982). "Trends and random walks in macroeconomic time series: Some evidence and implications," *Journal of Monetary Economics*, 10, 139-162.

Park, J. Y. and P. C. B. Phillips (1988). "Statistical inference in regressions with integrated processes: Part 1," *Econometric Theory*, 4, 468-497.

Park, J. Y. and P. C. B. Phillips (1989). "Statistical inference in regressions with integrated processes: Part 2," *Econometric Theory*, 5, 95-131.

Paulsen, J. (1984). "Order determination of multivariate autoregressive time series with unit roots," *Journal of Time Series Analysis*, 5, 115-127.

Phillips, P. C. B. (1991). "To criticize the critics: An objective Bayesian analysis of stochastic trends," *Journal of Applied Econometrics*, 6(4), 333-364.

Phillips, P. C. B. and W. Ploberger (1991). "Time series modeling with a Bayesian frame of reference: I. Concepts and illustrations," Cowles Foundation Discussion Paper No. 980.

Phillips, P. C. B. and W. Ploberger (1992a). "Time series modeling with a Bayesian frame of reference: II. General theory and applications," in preparation.

Phillips, P. C. B. and W. Ploberger (1992b). Bayesian model selection and prediction with empirical illustrations," in preparation.

Pötscher, B. M. (1989). "Model selection under nonstationarity: Autoregressive models and stochastic linear regression models," *Annals of Statistics*, 17, 1257-1274.

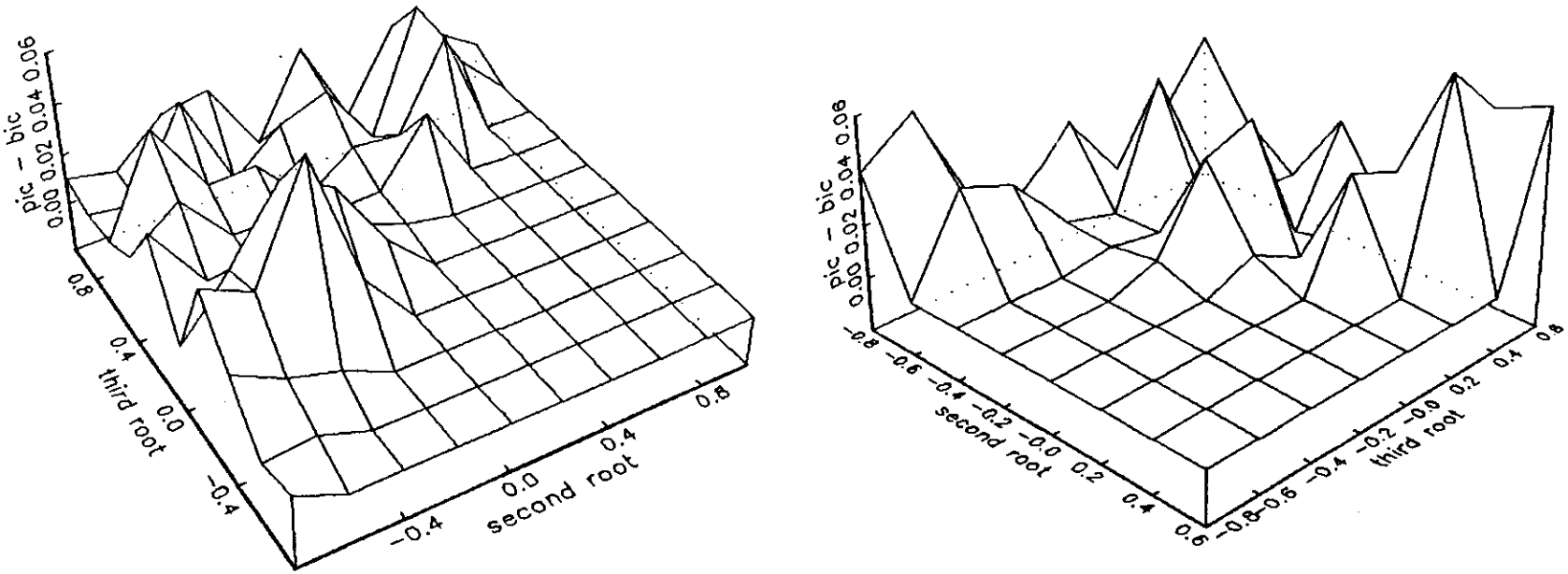
Rissanen, J. (1978). "Modeling by shortest data description," *Automatica*, 14, 465-471.

Schwarz, G. (1978). "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.

Tsay, R. S. (1984). "Order selection in nonstationary autoregressive models," *Annals of Statistics*, 12, 1425-1433.

Wei, C. Z. (1992). "On predictive least squares principles," *Annals of Statistics*, 20, 1-42.

Difference in P(correct model choice) between PIC & BIC



(a) first root = 1.00

(b) first root = 0.60

Figure 1: PIC − BIC Difference