Cowles Foundation Paper 798

JOURNAL OF APPLIED ECONOMETRICS, VOL. 6, 333–364 (1991)

TO CRITICIZE THE CRITICS: AN OBJECTIVE BAYESIAN ANALYSIS OF STOCHASTIC TRENDS

P. C. B. PHILLIPS

Cowles Foundation for Research in Economics, Yale University, USA

SUMMARY[1]

In two recent articles, Sims (1988) and Sims and Uhlig (1988/1991) question the value of much of the ongoing literature on unit roots and stochastic trends. They characterize the seeds of this literature as 'sterile ideas', the application of nonstationary limit theory as 'wrongheaded and unenlightening', and the use of classical methods of inference as 'unreasonable' and 'logically unsound'. They advocate in place of classical methods an explicit Bayesian approach to inference that utilizes a flat prior on the autoregressive coefficient. DeJong and Whiteman adopt a related Bayesian approach in a group of papers (1989a,b,c) that seek to re-evaluate the empirical evidence from historical economic time series. Their results appear to be conclusive in turning around the earlier, influential conclusions of Nelson and Plosser (1982) that most aggregate economic time series have stochastic trends. So far these criticisms of unit root econometrics have gone unanswered; the assertions about the impropriety of classical methods and the superiority of flat prior Bayesian methods have been unchallenged; and the empirical re-evaluation of evidence in support of stochastic trends has been left without comment.

This paper breaks that silence and offers a new perspective. We challenge the methods, the assertions, and the conclusions of these articles on the Bayesian analysis of unit roots. Our approach is also Bayesian but we employ what are known in the statistical literature as objective ignorance priors in our analysis. These are developed in the paper to accommodate explicitly time series models in which no stationarity assumption is made. Ignorance priors are intended to represent a state of ignorance about the value of a parameter and in many models are very different from flat priors. We demonstrate that in time series models flat priors do not represent ignorance but are actually informative (*sic*) precisely because they neglect generically available information about how autoregressive coefficients influence observed time series characteristics. Contrary to their apparent intent, flat priors unwittingly bias inferences towards stationary and i.i.d. alternatives where they do represent ignorance, as in the linear regression model. This bias helps to explain the outcome of the simulation experiments in Sims and Uhlig and some of the empirical results of DeJong and Whiteman.

Under both flat priors and ignorance priors this paper derives posterior distributions for the parameters in autoregressive models with a deterministic trend and an arbitrary number of lags. Marginal posterior distributions are obtained by using the Laplace approximation for multivariate integrals along the lines suggested by the author (Phillips, 1983) in some earlier work. The bias towards stationary models that arises from the use of flat priors is shown in our simulations to be substantial; and we conclude that it is unacceptably large in models with a fitted deterministic trend, for which the expected posterior probability of a stochastic trend is found to be negligible even though the true data generating mechanism has a unit root. Under ignorance priors, Bayesian inference is shown to accord more closely with the results of classical methods. An interesting outcome of our simulations and our empirical work is the bimodal Bayesian posterior, which demonstrates that Bayesian confidence sets can be disjoint, just like classical confidence intervals that are based on asymptotic theory. The paper concludes with an empirical

[1] All citations in this Summary are from Sims (1988) and Sims and Uhlig (1988). They are repeated in full in the text of this paper, where their precise locations in the cited articles are given.

0883–7252/91/040333–32$16.00

Received August 1990

Revised December 1990

application of our Bayesian methodology to the Nelson–Plosser series. Seven of the 14 series show evidence of stochastic trends under ignorance priors, whereas under flat priors on the coefficients all but three of the series appear trend stationary. The latter result corresponds closely with the conclusion reached by DeJong and Whiteman (1989b) (based on *truncated* flat priors). We argue that the DeJong–Whiteman inferences are biased towards trend stationarity through the use of flat priors on the autoregressive coefficients, and that their inferences for some of the series (especially stock prices) are fragile (i.e. not robust) not only to the prior but also to the lag length chosen in the time series specification.

Readers, even mature readers, are attracted to a writer who is quite sure of himself (T. S. Eliot, 1961).

1. INTRODUCTION

Since the influential empirical article by Nelson and Plosser (1982) on trends and random walks in economic time series there has been an explosion of interest in the econometrics of unit roots and stochastic trends. This interest has brought together theory and application in a way that is unusually productive for a new field. Together with subsequent developments on cointegration, the theory has given rise to a large and growing volume of empirical research. Economists who do empirical work with macroeconomic time series have been excited by the knowledge that regression with nonstationary time series is better understood and, as a result, they have become more confident in the interpretation of their empirical results. The excitement is understandable in light of the fact that as little as six years ago there was no theory of regression applicable to nonstationary series. The recent article by Stock and Watson (1988) well illustrates the empirical relevance of the new regression theory for nonstationary series, and the many ways in which it can assist our understanding of economic time series. However, intellectual acceptance of the methods of unit root econometrics has not been universal, and a wave of scepticism of the field, criticism of its methodology and re-evaluation of its empirical findings based on an alternative Bayesian methodology has recently appeared.

Initiating this wave of criticism in a highly sceptical essay, Sims (1988) put forward the view that classical inference procedures are misleading in models with unit roots, and argued that Bayesian methods are simpler to use, lead to more reasonable inferences and are largely unaffected by the presence of unit roots. Classical procedures, he suggested, are to be mistrusted

> *precisely because they do differ substantially from Bayesian procedures in this context* (Sims, 1988, p. 474). (S_1)

(This and all subsequent citations from Sims (1988) will be labelled in the form (S_i), $i = 1, \ldots, n$.)

In a sequel to that article, and using Monte Carlo simulations, Sims and Uhlig (1988/1991) provided a visual helicopter view of the joint probability density of the unknown autoregressive coefficient ρ and its least-squares estimate $\hat{\rho}$ in the simple AR(1)

$$y_t = \rho y_{t-1} + u_t \qquad (t = 1, \ldots, T) \tag{1}$$

with (u_t) i.i.d. $N(0, \sigma^2)$. Computed under a flat prior for ρ and with $\sigma^2 = 1$, their figures show the symmetric conditional distribution of $\rho \mid \hat{\rho}$ and the asymmetric conditional distribution of $\hat{\rho} \mid \rho$, thereby illustrating the operational differences between Bayesian and classical inference procedures in this context. They also compute the prior that would be implied by treating classical significance levels as if they were Bayesian posterior probabilities. They conclude as follows:

> *Use of classical statistical tests as measures of the plausibility of hypotheses is logically unsound. We have shown that in the case of a simple time series model with a unit root it amounts to acting as if one had a stronger prior belief in a root at or above one, the closer to one is the estimated value $\hat{\rho}$ of the root* (Sims and Uhlig, 1988, pp. 8–9). (SU_1)

Both extracts (S_1) and (SU_1) well represent the scepticism about classical procedures of inference that is the central message of these two papers. The ring of confidence with which they are written is certain to attract other researchers, as our header by T. S. Eliot suggests, and it will do so almost irrespective of the merits of the case.

Adopting a related Bayesian approach, DeJong and Whiteman have recently launched a series of independent empirical investigations (1989a,b,c) which seek to re-evaluate by Bayesian methods the evidence in support of unit roots and stochastic trends in the macroeconomic time series. Their philosophy marries well with that of Sims, their methodology follows the example of Geweke (1988) in the use of flat priors for the time series coefficients, and their empirical results appears to be conclusive. In reconsidering the historical time series studied originally by Nelson and Plosser (1982), DeJong and Whiteman (1989b) discover that trend stationarity is much more likely in terms of the Bayesian posteriors than difference stationarity. Only when zero prior probability is attached to trend stationary alternatives, they argue, will the AR representation of most macroeconomic time series appear to contain a unit root. They sum up their empirical re-evaluation by telling us that:

> *the death of trend stationarity appears to have been greatly exaggerated* (DeJong and Whiteman, 1989b, p. 13).

The purpose of the present paper is simple. We seek to challenge the methods, the assertions, and the conclusions of these articles on the Bayesian analysis of unit roots, and we offer an alternative methodology in its place. Our own approach is also explicitly Bayesian. But we differ by employing objective ignorance priors rather than flat priors in our analysis. As we shall explain below and more fully in Section 3, the epithets 'objective' and 'ignorance' are used advisedly in defining these priors, and they correspond to established usage in the statistical literature, although the precise forms for the priors that we develop here are new. Our analysis will help to illustrate how, in contrast to the thrust of (S_1), the Bayesian approach can lead just as easily to inferences that are compatible with those of classical procedures as it can to divergent inferences. This shows the fragility of Bayesian inferences about unit roots and stochastic trends to the specification of the prior. Moreover, objective Bayesian analysis reflects as much uncertainty about the data generating mechanisms as classical significance testing. Far from being 'logically unsound', as claimed in (SU_1), classical asymmetric sampling distributions are simply a manifestation of this uncertainty. The analogue of this phenomenon in objective Bayesian inference is, as we shall show, a bimodal posterior distribution of $\rho \,|\hat{\rho}$, which is a striking consequence of the use of ignorance priors in place of flat priors in the analysis.

The message of this study, like its purpose, is simple: when Bayesian and classical procedures lead to divergent conclusions we should seek first to find the answer in the prior rather than rush out to announce the failure of classical methods. What seems to have obscured this natural answer in the present case is the mistaken supposition that flat priors are uninformative and representative of ignorance. In a time series setting they certainly are not and, in consequence, they need to be used with more care and more qualifications in inference than we believe the articles cited above demonstrate.

The plan of the paper is as follows. In Section 2 we confront the scepticism articulated in Sims (1988) about the methodology of unit root econometrics and we deal *seriatim* with each of his criticisms. In every case we find his grounds for doubt to be unfounded. In our view his assertions about the impropriety of classical methods of inference are *ex cathedra*, unjustified and, in some cases that we make explicit, plain wrong. His claims about the superiority of Bayesian methods under flat priors are unwarranted. Indeed, we regard neither classical nor Bayesian approaches to be inherently 'unreasonable'. But, somewhat ironically in view of Sims' claims about its superiority, we show that the mechanical use of a flat prior Bayesian analysis is itself unreasonable because, contrary to apparent intent, such priors are informative in autoregressions and they unwittingly downweight the possibility of unit root and explosive alternatives. Section 3 introduces an alternative Bayesian approach based on ignorance priors that seek to represent the notion that a parameter is completely unknown. Such an approach is said to be objective, as distinct from subjective, Bayesian and it goes back to early work by Jeffreys (1946) and Perks (1947). We develop ignorance priors for the autoregressive coefficient ρ in model (1) and similar autoregressive models with trends and more general transient dynamics. The joint posterior for ρ and the other parameters is given under a Gaussian likelihood and the marginal posterior for ρ is obtained analytically by using a Laplace approximation to reduce the multidimensional integral. Sections 3.2–3.4 report simulations which evaluate the new procedure against the flat prior Bayesian approach. The bias towards stationary and trend stationary alternatives in posteriors obtained from flat priors is found to be substantial in every case. Indeed, in a model such as (1) with a fitted trend, a flat prior on ρ and $T = 50$, we would expect, on average, *when the true data-generating mechanism has a unit root* to find the posterior probability of nonstationarity, viz. $P(\rho \geqslant 1)$, to be less than 5 per cent. This degree of bias seems unacceptable by most standards. In this assessment we use the frequentist terminology 'bias' deliberately to describe Bayesian posteriors which are sufficiently mislocated that the true generating mechanism (here, one that involves stochastic nonstationarity) is made to appear unlikely in posterior probability calculations. Section 4 reports the results of an empirical illustration of our methods to the Nelson–Plosser time series.

2. SCEPTICISM CONFRONTED

In his (1988) paper Sims questions the value of much of the ongoing work on unit root inference in econometrics and claims that the seeds of this work 'are essentially sterile ideas' (p. 463). If one were to interpret sterility literally as an incapacity to produce offspring, then the fecundity of the research in the field would itself belie that claim. Notwithstanding this irony, several explicit 'grounds for doubt' about the value of classical inferential procedures and arguments in support of the assertion about sterility are given by Sims, although the arguments that are offered are only brief and are largely nontechnical. The central argument is the divergence of Bayesian and classical inference expressed in (S_1) and this we shall address in Sections 3 and 4. However, since we wish to be complete in this critique, since some of the attendant issues are themselves of interest, and since Sims's prescriptions and scepticisms are being taken seriously by other researchers, we shall look here explicitly at the stated grounds for doubt. We shall deal with them individually and in the order in which they appear in the cited paper.

(a) Tenuous Connections Between the Unit Root Hypothesis and Economic Theory

The efficient markets hypothesis for asset prices is one of the main behavioural economic theories that lead to models with unit roots. Sims argues that this model is at best just an

approximation that applies for small time intervals. Similarly, to present his case here, Hall's (1978) martingale model for consumption strictly applies only under rigid conditions on utility and under assumptions like constant real interest rates which hardly seem tenable except over short time periods. Likewise, models that incorporate technological change via stochastic processes with unit roots have only tenuous connections with economic theory.

There is validity in each of these objections. Yet similar objections of specificity and approximation can be raised against most economic theory, more especially macroeconomic theory that is based on representative agent paradigms. Models like the permanent income hypothesis and the efficient markets hypothesis, it should be remembered, are powerful in their predictions and useful in terms of their interpretative content precisely because of their simplicity. Moreover, in spite of a long history of objections, these models, as distinct from innumerable others, have survived and evolved as theoretical constructs. The efficient markets hypothesis, in particular, has continued to perform well empirically against all competitors. Few theory models can claim a comparable degree of success and longevity. Were it not for these empirical successes, and for the underpinning in efficient markets theory, it would surely be unlikely that a root of unity would be selected as the leading prior mean in so many Bayesian VAR exercises.

To the extent that both behavioural and empirical models are approximations to an evolving time series reality we can expect that any model will retain its relevance only over finite spans of data. As more data are brought to bear, it is common to find that the variance of the prediction error increases linearly over time. In other words, the superposition of new shocks over time leads to stochastic drift away from a given model and its best predictions. Such stochastic drift constitutes strong empirical evidence in favour of the unit root hypothesis. It can be incorporated by direct reasoning in modelling as in the efficient markets theory, or indirectly as in real business cycle models where the ultimate engine of change in the economy is taken to be the demographic and technological supply-side shocks that affect the economy's productive capacity. In either case the effect is the same and, in consequence, the unit root hypothesis is about as well connected to the behavioural economic theory that appears in time series models as any other justifiable empirical feature of those models.

Some of the latest perspectives in macroeconomic thinking have actually strengthened the links between unit roots and behavioural economic theory. In particular, work by Durlauf (1989, 1990) has shown that coordination failure models with incomplete markets and multiple equilibria can generate unit roots from shocks that enter the system period by period, irrespective of their origin in demand or supply-side disturbances. Moreover, unit roots can occur in these models even when technical change is deterministic.

Thus, the Sims objections to unit roots on this ground have some validity as generic criticisms of economic theory and, in our view, they are comparable to earlier criticisms voiced in Sims (1982) of representative agent rational expectations modeling as a 'revolution [that] itself has had its excesses, destroying or discarding much that was of value in the name of utopian ideology' (p. 107). However, Sims' objections ignore the longevity and the successes of the efficient markets theory, they overlook the importance of sophisticated simplicity in modeling (as argued, for instance, by Friedman, 1953 and Zellner, 1988), they fail to take into account the latest thinking in macroeconomic modelling, and they are inconsistent with the pervasive use of unit root priors in VAR empirical models.

(b) Mistaken Perspectives on the Effects of Unit Roots on Classical Inference

It is by now well understood that the presence of unit roots does affect asymptotic distribution theory and classical procedures of inference. Indeed, much of the ongoing literature has been

concerned with the many different consequences of this fact. Sims recognizes this but then tells us that:

> *The attempt to apply asymptotic distribution theory allowing for nonstationarity has been in most instances wrongheaded and unenlightening* (Sims, 1988, p. 464). (S_2)

No examples or citations to support this view of community-wide bungling are given. The reader is instead referred to Sims, Stock, and Watson (1990) for a demonstration of the fact that in linear VARs conventional $\sqrt{T}$ normal asymptotics apply, albeit with some degeneracies depending on the number of unit roots in the system. This description of lowest-level normal asymptotics is perfectly accurate when there are stationary or cointegrated regressors, and it applies much more generally to misspecified systems, as shown by Park and Phillips (1989). However, this is far from being the whole story. Unhappily, the degenerate $\sqrt{T}$ normal asymptotics have led many to conclude mistakenly that conventional asymptotic tests apply without modification in nonstationary models. Indeed, Sims himself errs on this point when he concludes that:

> *any hypothesis which can be tested after the model is transformed* [*to stationary form*], *can be tested with exactly the same distribution theory using the untransformed model. There is no justification for preliminary differencing or application of cointegration transformations in the belief that these steps are necessary to allow use of the usual statistical tests* (Sims, 1988, p. 465; my insertion in square brackets, for purposes of clarification). (S_3)

A major counterexample to this statement is given in my paper (Phillips, 1988/1991) on optimal inference in cointegrated systems. As argued there, linear VARs in levels or log levels implicitly estimate whatever roots, including unit roots, there may be in the system. This means that estimates of any cointegrating relationships in the system have a limit theory that depends on the limit distributions that apply for the estimated unit roots. Moreover, as explained in my (1988/1991) paper, estimated cointegrating vectors obtained from VARs in levels suffer from simultaneous equations bias, a somewhat ironic outcome given the arguments put forward a decade ago by Sims (1980) for the use of VARs in place of simultaneous equations models. By contrast, when the model is transformed to its stationary error correction model (ECM) representation, these problems do not appear because the unit roots are no longer estimated when the model is in this format. Instead, estimates of cointegrating vectors from ECM formulations are optimal and follow a mixed normal limit theory. As a result, tests of hypotheses about the cointegration space can be conducted validly with usual asymptotic chi-squared criteria. This is not possible for the untransformed VAR in levels formulation. Similar arguments apply also to causality tests, although in the case of these tests there are major asymptotic problems even in ECM formulations. Thus, (S_3) is simply wrong on this point.

More generally, it is important to recognize that the likelihood ratio is not locally asymptotically normal (LAN) in the sense of LeCam (1960) when there are unit roots to be fitted. In fact, the likelihood ratio is not even locally asymptotically quadratic in this case, as shown in Proposition 4.1 in Phillips (1989). The reason is that the information (in the sense of R. A. Fisher) that is carried by the data about the unit root is both random and variable (i.e. sensitive to local departures from unity) and this uncertainty persists even in asymptotic samples. Appendix A provides more technical detail on this point. But when the model is transformed to stationary form the likelihood ratio is LAMN (i.e. locally asymptotically mixed normal, as in Jeganathan, 1980) and all of the inferential theory, including optimality, for LAMN families applies.

Thus, in direct contrast to the assertion (S_3) there is substantial justification in terms of asymptotic distribution theory and optimality theory for working with transformed specifications such as ECM formulations rather than untransformed VARs. If one takes into account that VARs in levels produce estimates of the cointegration space that suffer from simultaneous equations bias (Phillips, 1988/1991), causality tests that generally involve nonstandard limit theory and nuisance parameters (Sims, Stock, and Watson, 1990) and impulse response functions that are both arbitrary (Cooley and LeRoy, 1985) and very imprecise (Runkle, 1987), there would seem to be little justification for using VARs empirically, even if one's preferred modelling methodology is atheoretical.

(c) The Discontinuity in the Classical Asymptotic Theory at $\rho = 1$ Generates Confidence Regions of 'Disconcerting Topology'

The argument is as follows. If a fitted value $\hat{\rho} < 1$ with t-ratio $t_1(\hat{\rho}) = (\hat{\rho} - 1)/s_{\hat{\rho}}$ leads to acceptance of a unit root null under the unit root limit theory for $t(\hat{\rho})$ but rejection under conventional normal asymptotics, then classical confidence regions can be disconnected because of the exclusion of some values of ρ close to unity from the confidence set since the corresponding t-ratio $t_\rho(\hat{\rho}) = (\hat{\rho} - \rho)/s_{\hat{\rho}}$ would reject them. The phenomenon arises because the asymptotic critical values under a unit root null are further out in the left tail than those of a stationary null for ρ close to but less than unity. Sims finds this feature of the classical approach disconcerting, and argues that Bayesian inference encounters no such difficulties because

> *The likelihood, and hence the posterior p.d.f. for a flat prior, is Gaussian in shape regardless of whether or not there are unit (or even explosive) roots. This simple flat-prior Bayesian theory is both a more convenient and a logically sounder starting place for inference than classical hypothesis testing.* (S_4)

This is a strong and confident assertion. Yet the flat prior condition under which it is given is nowhere near as innocent as it appears, nor is the data conditioning that is part of the Bayesian analysis. In fact, Bayesian inference in time series models under flat priors for the coefficients is formally identical to that of the linear regression model in which the regressors are fixed and nonrandom. No consideration is given to the time series nature of the data either in setting the prior or in conditioning on sample moments. Of course, Bayesian inference typically pays little attention to the sample space, gives maximum attention to the parameter space, and always proceeds by conditioning on the observed data.[2] Moreover, flat priors are convenient to use, they have established precedent in earlier work (e.g. Zellner, 1971) and in the normal linear regression model they are well known to lead to Bayesian confidence sets that are equivalent to the corresponding sampling theory (e.g. Malinvaud, 1980, pp. 239–240).

Why is the situation so grossly different in a time series setting? The reason is that in the normal linear regression model the coefficients influence only the mean of the data and conditioning on fixed regressors is innocuous. In a time series model, on the other hand, the coefficients influence the mean, the variance, and the entire autocorrelation structure of the data and conditioning on the random sample moment matrices of time series data is not innocuous. Indeed, the values of the coefficients in time series models actually influence the

[2] This characterization of Bayesian procedures is by no means simply a personal view. It is recurrent in many discussions of Bayesian theory For a recent example the reader is referred to the discussion of Lindley (1990) and, in particular, to the comments of Lehmann (1990).

amount of information that is carried in the data and its sample moments. This is especially true of statistics like the sample variance of the regressor in an AR(1) like (1). To condition on such a sample variance, as Sims does implicitly in (S_4) when he tells us that the likelihood is Gaussian in shape, is to ignore the information that it carries about the coefficient ρ. Worse than this, the claim of a 'Gaussian likelihood' is made possible only by using the sample variance of the regressor to determine the units of measurement of departures of ρ from $\hat{\rho}$, and this is tantamount to the use of a variable yardstick, one that relies on ρ itself (as the asymptotics certainly make clear). I hasten to add that there is nothing inherently wrong with this particular conditioning device that changes the units of measurement, provided it is clearly understood that the frame of reference or geometry of the problem has been changed. In fact, if one wants to look at the data in this new frame of reference then one can easily do so in a classical frequentist approach. Indeed, one can go much further than the Sims statement (S_4) and demonstrate that fixing the amount of information in the data is an effective device for normalizing the distribution of an estimator such as $\hat{\rho}$. None of this means that either Bayes or classical theory is the 'more convenient and logically sounder starting place for inference'. But it does show that with an appropriate frame of reference the two approaches to inference are certainly much closer than (S_4) implies.

In time series models, flat priors ignore the way in which the coefficients influence the amount of information contained in the sample. A flat prior on ρ in model (1), for instance, deems that it is equally likely for ρ to be in the two intervals [0·50, 0·60] and [0·95, 1·05]. Yet this prior ignores what we already know about the effects of ρ in these different intervals on sample behaviour. Typical trajectories, responses to shocks, and changes in initial conditions are all very different for ρ in these two intervals, and constitute prior knowledge based on our understanding of the AR(1) model. In this context, flat priors on the autoregressive coefficient cannot represent ignorance in any meaningful sense. In fact, the next section will demonstrate, they are highly informative, they lead to inferences about the presence of stochastic trends and unit roots that are often severely biased against these possibilities, and they can give a misleading impression of precision in inferences.

One way to give due consideration to the time-series nature of the data is to use an objective ignorance prior such as that suggested originally by Jeffreys (1946). As illustrated in Sections 3 and 4, under such a prior the Bayesian posteriors for the autoregressive coefficient $\rho \mid \hat{\rho}$ in models like (1) are then frequently bimodal and lead to disjoint confidence sets, just as those based on classical sampling theory asymptotics. This is a possibility not recognized by Sims. Far from being 'logically unsound', we find that classical procedures lead to inferences that are often close to their Bayesian counterparts under appropriate ignorance priors. There is no fatal flaw in either approach to inference, simply human error in accepting conclusions too readily from fragile and informative priors. The uncertainty about the data generating mechanism that manifests itself in disjoint confidence sets and low power in unit root tests is itself present in Bayesian inference when due allowance is made for the time series nature of the data in the construction of an uninformative prior. Moreover, the fragility of Bayesian inferences to the specification of the prior should itself be taken as a signal of this uncertainty, as indeed it is by some Bayesians such as Leamer (1983, 1988).

(d) The Classical Approach Ignores Useful Evidence Against $\rho = 1$

Sims puts forward the following explanation of his position:

One of the unreasonable aspects of the classical approach to this problem is that likelihood ratio tests make no use of our knowledge that a large σ_ρ in a large sample is evidence against $\rho = 1$ even if the t-statistic for $\rho = 1$ is fairly small (Sims, 1988, p. 471). (S_5)

Here, $\sigma_\rho = \sigma\{\Sigma y_{t-1}^2\}^{-1/2}$ is a 'standard error' for $\hat{\rho}$. Its asymptotic behaviour depends on the value of ρ. Thus, when $|\rho| < 1$ we have $\sigma_\rho = O_p(T^{-1/2})$ and when $\rho = 1$ we have $\sigma_\rho = O_p(T^{-1})$, leading us to expect smaller 'standard errors' for $\hat{\rho}$ in large samples in models with a unit root. Thus, we agree with the latter part of (S_5) describing our knowledge about σ_ρ. But we dispute the claim in (S_5) concerning the unreasonable aspect of the classical approach. Indeed, it is the Bayesian approach under flat priors, not classical methods, that ignore this generic information about σ_ρ in time series models like (1). We make the following points.

1. Under the null hypothesis that $\rho = 1$ we may estimate σ_ρ^2 by $\hat{\sigma}_\rho^2$ where

$$T\hat{\sigma}_\rho^2 = \Sigma(y_t - y_{t-1})^2 / \Sigma y_{t-1}^2. \tag{2}$$

   This statistic is the Von Neumann ratio of the Gaussian random walk. Its use as a statistic for testing for the presence of a unit root and for testing the specification of a regression equation in levels or differences (where regression residuals are employed in place of y_t in (2)) was considered by Dickey and Fuller (1981), Berenblut and Webb (1973), Sargan (1979), and Sargan and Bhargava (1983). Indeed, the statistic may be interpreted as the likelihood ratio test of the null of serial dependence against the alternative of a random walk and, as discussed by Sargan and Bhargava (1983), it is known to be a most powerful test in a neighbourhood of the alternative. A closely related version of this statistic has recently been obtained as an LM test for a unit root in Schmidt and Phillips (1989). Thus, to argue as in (S_5) that the classical approach ignores evidence based on σ_ρ, is simply to fly in the face of the facts.
2. Sims claims that 'when $\rho = 1$, σ_ρ behaves asymptotically like a constant times $1/T$' (1988, p. 470). In fact, when $\rho = 1$, σ_ρ behaves like a *random variable* times $1/T$. The difference is nontrivial and has important consequences. First, it causes a breakdown in the local asymptotic quadratic property of the likelihood, as discussed under (b) above and in Appendix A. Second, since the limit random variable carries information about ρ as seen from equation (A1) of Appendix A, one might well expect that conditioning on the sample moment $T^{-2}\Sigma y_{t-1}^2$ would involve a loss of information. Actually, Bayesian conditioning on the data does just this under flat priors, i.e. it treats time series data like data from a linear model with fixed regressors whereas, depending on the value of ρ, the sample moments of the data may have radically different behaviour. It is for this very reason that flat priors in time-series models are informative. They suggest that we believe all values of ρ to be equally likely when, in fact, we know that large values of ρ are much more likely when scale parameters or standard errors like σ_ρ are very small. The ignorance priors we use in the following section explicitly take this balance into account. Priors like flat priors do not and thereby, are unwittingly informative in time series models.

To sum up, we submit that Sims errs on two counts in (S_5): first, many classical statistics take the scale effects σ_ρ into account and some, like the Von Neumann ratio (3), are constructed directly from it; second neither classical nor Bayesian approaches are inherently 'unreasonable', but, somewhat ironically in view of the claim in (S_5), the mechanical use of

flat priors in time series models is unreasonable because, contrary to apparent intent, such priors are informative and can thereby seriously and unwittingly bias inferences. Section 3 will give examples.

3. OBJECTIVE IGNORANCE PRIORS AND UNIT ROOTS

3.1. The Justification of Ignorance Priors

In a subjectivist approach to Bayesian inference the role of a prior distribution is to represent the degree of subjective belief of the person who makes the inference. Partly because of the difficulties associated with prior elicitation, and partly because there is a need in many applications to proceed under conditions that approximate ignorance, many Bayesian writers have sought to establish an objective basis for the choice of the prior. In an objective theory, the prior seeks to represent the notion that a parameter is completely unknown, thereby giving rise to the term 'ignorance prior'.

Jeffreys (1946) was the first to suggest a method for inducing ignorance priors in a given probability model. Earlier researchers had followed Bayes and assumed that ignorance could be represented by a uniform distribution (i.e. a diffuse or flat prior) over the parameter space. Yet, as is now well known, flat priors on different versions of the parameter space yield different posteriors, i.e. the posterior is not invariant to 1:1 transformations of the parameter space. Jeffreys's idea was to base the selection of the objective prior on certain invariance properties of the family of probability densities $f(x|\theta)$, indexed by the parameter $\theta \in \Theta$, from which the data were drawn. The prior so selected would then inherit those invariance properties and thereby avoid any arbitrariness in the choice of parameters since it would assign the same prior probability to equivalent propositions (i.e. irrespective of their parameterization). If we set $I_{\theta\theta} = -E\{(\partial^2/\partial\theta\,\partial\theta')\log(f(x|\theta))\}$ then Jeffrey's general suggestion was the prior

$$\pi(\theta) \propto |I_{\theta\theta}|^{1/2} = J(\theta), \text{ say.} \tag{3}$$

This prior is invariant in the above-mentioned sense to smooth transformations of the parameters $\varphi = \varphi(\theta)$ because of the equivalence of the corresponding probability elements

$$|I_{\theta\theta}|^{1/2}\,d\theta = |I_{\varphi\varphi}|^{1/2}\,d\varphi, \tag{4}$$

(e.g. Jeffreys, 1961, p. 180); Zellner, 1971, p 48; Box and Tiao, 1973, pp. 41–46).

Hartigan (1964) showed that the Jeffreys prior (3) has other useful invariance properties of which the most important are its invariance under (i) smooth data transformations (e.g. changes in the units of measurement), (ii) restrictions in the parameter space, (iii) replication of the sample space, and (iv) replacement of the data by a sufficient set of statistics. Subsequently, Hartigan (1965) showed that (3) is an asymptotically unbiased prior distribution under a Jeffreys loss function in the sense that the prior density (3) minimizes the asymptotic bias of the corresponding Bayes estimator (i.e. the estimator that minimizes expected loss).

An alternative justification for the Jeffreys prior was given by Lindley (1961). Lindley argued that knowledge of θ means knowing $f(x|\theta)$, and that the amount by which θ differs from $\theta + \delta(\theta)$ on some mesh of size $\delta(\theta)$ can, in turn, be measured by how much $f(x|\theta)$ differs from $f(x|\theta + \delta(\theta))$. Using Shannon's information criterion as the metric for this distance between the densities, and assigning a uniform prior on the interval $[\theta, \theta + \delta(\theta)]$ to represent ignorance (as distinct from the knowledge of θ), Lindley obtained the Jeffreys prior (3).

Another early suggestion for the generation of ignorance priors was made by Perks (1947), who argued that the prior distribution should reflect the anticipated asymptotic volume of confidence regions. Under general regularity conditions the confidence region around θ has volume that is asymptotically proportional to $J(\theta)^{-1}$. So if θ_0 is the true value we anticipate a tight confidence region near θ_0 if $J(\theta_0)$ is large. The Jeffreys prior (4) assigns a density to θ that reflects this expectation. Welch and Peers (1963) made this confidence region argument more explicit by showing that, asymptotically, one-sided Bayes confidence sets generated from Jeffreys prior are closer to classical confidence intervals than those of any other prior.

As far as our own application to time series models is concerned, the Perks justification of (3) is highly relevant. Thus, when $|\rho| \geqslant 1$ in model (1) we anticipate confidence regions for the true value ρ_0 to be tighter, indeed much tighter, than when $|\rho| < 1$. This expectation turns out to be properly represented in an ignorance prior on the autoregressive coefficient ρ. Thus, the true coefficient ρ_0 is completely unknown, but the ignorance prior still reflects the knowledge we have about the AR(1) model that were $|\rho|$ to be large, the data would be much more informative about ρ. This generic model characteristic that confidence sets will be tighter when $|\rho|$ is large is neglected in a flat prior. In treating all values of ρ as equally likely, the flat prior unwittingly carries information that downweights large values of ρ. In so doing, Bayesian inference under a flat prior on ρ will be distorted by information that will bias the posterior towards stationary alternatives. Simply put, flat priors are informative in time series models that permit nonstationarity and they inform by effectively downplaying the possibility of unit root and explosive alternatives. In time series models with deterministic trends it is therefore hardly surprising that Bayesian inference under flat priors strongly favours trend stationary alternatives.

Before leaving this introductory discussion it is worth remarking that, while the invariance properties of the Jeffreys prior have generally been viewed as desirable characteristics, there are features of Jeffreys priors that some Bayesians have found disagreeable. First, difficulties with regard to expected degrees of freedom have been encountered in the use of Jeffreys priors for multidimensional parameter spaces, leading Jeffreys himself (1961, p. 182) to propose modifications to the general rule (3). We shall comment further on this problem in our present application below. Second, the Jeffreys prior uses the model itself as the mechanism for generating prior probabilities. While this does provide an objective basis for Bayesian analysis, it nonetheless exposes the analysis to what inevitably must be rather arbitrary elements in the construction of the model, such as the time unit, choice of variable, number of regressors, treatment of initial conditions, and so on. However, as the author has recently argued elsewhere (Phillips, 1988), choices that underlie the construction of a model belong to the antecedent thinking in all good modelling where the investigator shapes his purpose. Both Bayesian and classical statistical methods are subject to the decisions made in this early stage of modelling. Even in a purely subjective Bayesian view, where probabilities express 'personal beliefs', the investigator must address and resolve many such modelling choices before he attempts to articulate his 'personal beliefs' in quantitative form.

3.2. A New Look at Bayesian Inference in the AR(1)

We start by considering the simple AR(1) model (1). Conditioning on the initial value y_0, the Gaussian likelihood follows from the density

$$f(y \mid \rho, \sigma, y_0) = (2\pi)^{-T/2}\sigma^{-T} \exp\{-(1/2)\sigma^{-2}\textstyle\sum_1^T (y_t - \rho y_{t-1})^2\}.$$

Assuming a flat prior for $(\rho, \log \sigma)$ leads to the usual purported 'uninformative' prior for (ρ, σ), *vis.*

$$\pi(\rho, \sigma) \propto 1/\sigma, \tag{5}$$

and Bayesian analysis of (1) under this prior is identical to that of the linear regression model. The joint posterior distribution is

$$p(\rho, \sigma \mid y, y_0) \propto \sigma^{-T-1} \exp\{-(1/2\sigma^2)[m(\hat{u}) + (\rho - \hat{\rho})^2 m(y)]\}, \tag{6}$$

where $\hat{\rho} = \Sigma y_t y_{t-1} / \Sigma y_{t-1}^2$, $m(y) = \Sigma y_{t-1}^2$, $m(\hat{u}) = \Sigma \hat{u}_t^2$ and $\hat{u}_t = y_t - \hat{\rho} y_{t-1}$. The marginal posteriors are:

$$p_F(\rho \mid y, y_0) \propto [m(\hat{u}) + (\rho - \hat{\rho})^2 m(y)]^{-T/2}, \tag{7}$$

$$p_F(\sigma \mid y, y_0) \propto \sigma^{-T} \exp\{-(1/2\sigma^2) m(\hat{u})\}. \tag{8}$$

Note that the marginal posterior for ρ is a univariate t_{T-1} distribution, ρ is symmetrically distributed about the OLS estimate $\hat{\rho}$ and the variance of ρ is $m(\hat{u})/(T-3)m(y)$, which decreases as $m(y)$ increases.

Thornber (1967) and Zellner (1971, Ch. VII) both used this framework and emphasized its applicability for stationary and nonstationary cases. Geweke (1988) used the same approach in a cross-country applied study but used a restricted domain in addition to the flat prior. Sims (1988) and Sims and Uhlig (1988/1991) also use this framework, although in the latter paper the model is even simpler because σ is assumed to be known for computational convenience. Schotman and van Dijk (1991) employ a similar approach in studying real exchange rate data. However, since their objective is to perform a posterior odds analysis of the unit root hypothesis, they modify (5) by truncating the domain over which ρ has a flat prior to a proper subset of the stationary interval and they assign a discrete prior probability mass to $\rho = 1$ (values of ρ in the explosive range being excluded). In all of these past studies, only Thornber and Zellner mention the possibility of a Jeffreys prior and they, along with Box and Jenkins (1976), and Jeffreys (1961, p. 187) in his original analysis, confine their attention to the stationary case.

In place of (5) we shall now consider a Jeffreys prior under which there is no stationarity assumption. As above, we will work conditional on the initialization y_0. Setting $\theta = (\rho, \sigma)$ we find, after a little calculation, that

$$I_{\theta\theta} = \begin{bmatrix} I_{\rho\rho} & 0 \\ 0 & I_{\sigma\sigma} \end{bmatrix},$$

with

$$I_{\rho\rho} = \begin{cases} \dfrac{T}{1-\rho^2} - \dfrac{1}{1-\rho^2}\dfrac{1-\rho^{2T}}{1-\rho^2} + \left(\dfrac{y_0}{\sigma}\right)^2 \dfrac{1-\rho^{2T}}{1-\rho^2}, & \rho \neq 1 \\ \dfrac{T(T-1)}{2} + T\left(\dfrac{y_0}{\sigma}\right)^2, & \rho = 1 \end{cases}$$

and

$$I_{\sigma\sigma} = 2T/\sigma^2.$$

The Jeffreys prior (4) is therefore given by

$$\pi(\rho, \sigma) \propto (1/\sigma) I_{\rho\rho}^{1/2}, \tag{9}$$

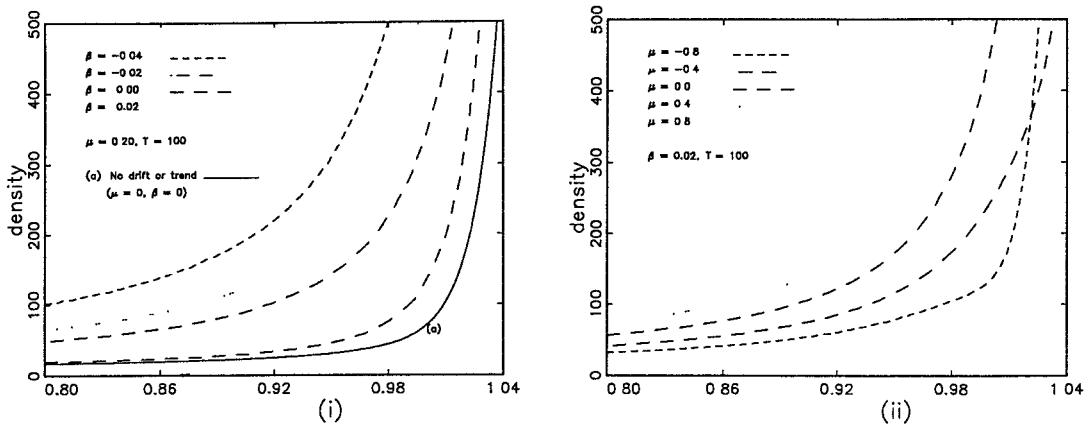

Figure 1. (i) & (ii) Ignorance priors for ρ

which is continuous in ρ for $-\infty < \rho < \infty$. Observe that this prior depends on y_0, which is the given initialization of the model, and the sample size T. Thus, just as the Jeffreys prior recognizes the information content of the sample variance of the regressor in this model, it also recognizes that the information content will grow as T increases and at a geometric rate when $\rho > 1$. The prior is graphed and displayed as curve (a) in Figure 1(i) for the case $y_0 = 0$, $T = 100$ and $\sigma = 1$; the log density is graphed as curve (a) in Figure 1(iii) and shows the density over a wider range of ρ values. Figure 1(i) shows how the prior increases slowly to the value $\{T(T-1)/2\}^{1/2}$ at $\rho = 1$ and then increases exponentially at the rate $O(\rho^{T-2})$ for $\rho > 1$. The higher density for $\rho > 1$ reflects the prior knowledge we always have from the model that when the true value of the autoregressive coefficient $\rho_0 > 1$ the data will carry more information about ρ_0. Aside from carrying this generic feature of the model, the prior is uninformative about ρ. As discussed in the preceding section, a flat prior on ρ is informative precisely because it neglects this generic characteristic of the model and the time series nature of the data.

The shape of the prior (9) as a function of ρ sheds light on the simulation exercise performed in Sims and Uhlig (1988/1991) whose outcome is summarized in the extract (SU_1). The implicit priors computed by Sims and Uhlig are purported to represent the prior under which classical p-values would correspond to Bayesian posterior probabilities conditional on $\hat{\rho}$. Although there is erratic sampling behaviour in the priors they compute, although their calculations are truncated just beyond unity and although, as they put it, their approach is

> *not formally justified by either a Bayesian or a classical argument* (Sims and Uhlig, 1988, p. 2), (SU_2)

it is apparent that their simulation results (Figures 8 and 9, in Sims and Uhlig, 1988) provide a very crude prior whose shape is not dissimilar to the Jeffreys prior (9), at least over the restricted domain they consider. Sims and Uhlig take the shape of their imputed prior as strong evidence of the unreasonableness of classical significance testing. Their assessment is based on comparison with a flat prior which they mistakenly regard as uninformative, and on the proposition that, were the ρ values truly uniformly distributed,

> *Everyone should agree that, on observing $\hat{\rho} = 1$, our uncertainty about ρ is symmetric about $\rho = 1$* (Sims and Uhlig, 1988, p. 6). (SU_3)

However, as our calculations below show, posteriors computed under the Jeffreys prior are not

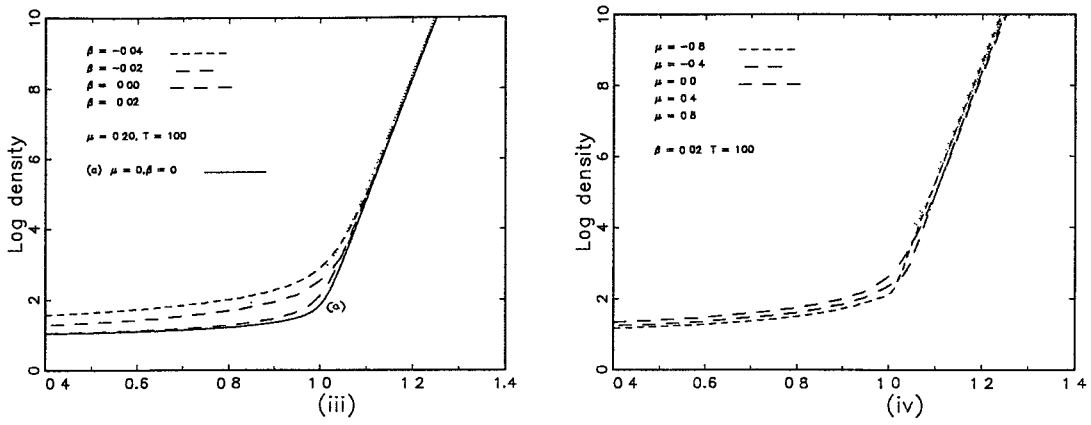

Figure 1. (iii) & (iv) Ignorance priors for ρ

symmetric, especially for values of $\hat{\rho}$ in the interval $\hat{\rho} \leqslant 1$. Note that a Jeffreys prior, even in the Sims–Uhlig hypothetical experiment, is still a very reasonable choice of prior, since once ρ is drawn we would still expect the data to be more informative about ρ the greater ρ is. (We cannot reasonably expect an investigator to be given the true distribution of ρ, for if this were available, there would be no point in collecting and using the data from a single trial. In a classical setting that would be equivalent to giving the investigator the true parameter value ρ_0 and then being surprised that he estimated ρ.) Thus, we see no reason to accept the proposition (SU_3), and we are surprised that it should be put forward as a universal belief. In our view the proposition arises from an intuition that comes from treating time-series models such as (1) like the linear regression model where flat priors on the coefficients have good properties.

Under the Jeffreys prior (9), the joint posterior is

$$p(\rho, \sigma \mid y, y_0) \propto \sigma^{-T-1} \exp\{-(1/2\sigma^2)[m(\hat{u}) + (\rho - \hat{\rho})^2 m(y)]\} I_{\rho\rho}^{1/2}, \tag{10}$$

and integration over σ gives the following marginal posterior for ρ when $y_0 = 0$

$$p_J(\rho \mid y) = \int_0^\infty p(\rho, \sigma \mid y, y_0 = 0)\, d\sigma \propto I_{\rho\rho}^{1/2} [m(\hat{u}) + (\rho - \hat{\rho})^2 m(y)]^{-T/2}. \tag{11}$$

Using the methods of Section 3.3 below it can be shown that (11) is an asymptotic approximation to the marginal posterior for ρ when $y_0 \neq 0$. Thus, y_0 will not substantially affect the posterior unless it is very large.

The marginal posterior (11) has a shape that can be very different from that of (7). Its main properties are discussed in the comments that now follow.

1. $p_J(\rho \mid y)$ has Pareto tails of order $O(|\rho|^{-2})$ as $|\rho| \to \infty$. Thus, upon standardization, (11) is a proper density. But its tails are like those of a Cauchy distribution and it has no finite integer moments.
2. Unlike (7), the density (11) is not symmetric about $\hat{\rho}$. It has one mode close to $\hat{\rho}$ and, depending on the values of $m(\hat{u})$ and $m(y)$, it often has a significant second mode for some $|\rho| > 1$.
3. When the true coefficient $\rho_0 = 1$ in (1), the asymptotic behaviour of the density based on

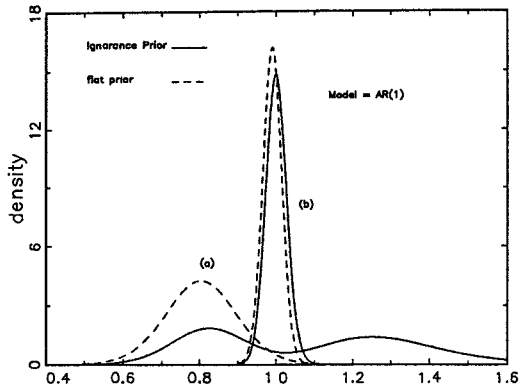

Figure 2. Posterior densities for ρ

(11) depends on that of

$$I_{\rho\rho}^{1/2}\left[1 + T(\rho - \hat{\rho})^2 \int_0^1 W^2\right]^{-T/2}$$

which we see to be of $O(T)$ for $\rho = 1$, of $O(T^{-(T-1)/2})$ for $0 < \rho < 1$ and of $O(T^{-T/2})$ for $\rho > 1$. Thus, Bayes estimators that are based on (11) are consistent but at a faster rate for $\rho > 1$ than for $\rho < 1$.

4. Figure 2 illustrates typical shapes of the posterior (11) for data generated from a random walk with initialization $y_0 = 0$ and $T = 50$. The figure graphs the normalized posterior densities based on a Jeffreys prior against those based on a flat prior. The figure displays the posteriors for two different data sets simulated from the model (1) with $\rho = 1$, and these are designated (a) and (b), respectively. The results are chosen because they are representative of the typical posterior shapes that emerge from a large number of simulations. The sample data characteristics for the two figures are as shown in Table I.

The flat prior posteriors (hereafter, F-posteriors) have symmetric bell shapes centered on the regression estimate $\hat{\rho}$. In the case of the curves designated (a), the estimated regression coefficient $\hat{\rho} = 0{\cdot}804$ is low and the F-posterior is so seriously biased downwards that the posterior probability, of $\rho \geqslant 1$, i.e. $P_F(\rho \geqslant 1) = 0{\cdot}02$, is negligible. By contrast, the Jeffreys prior posterior (hereafter, J-posterior) is bimodal in case (a). The principal mode is located slightly to the right of $\hat{\rho}$ and there is a second mode around the value $1{\cdot}25$. The posterior probability $P_J(\rho \geqslant 1) = 0{\cdot}54$ is appreciable. Thus, while the F-posterior effectively rules out a true ρ of unity, the J-posterior indicates considerable uncertainty about ρ and a true ρ of unity would definitely not be ruled out. Note that because of the bimodality of the J-posterior, Bayes confidence sets of shortest length would be disjoint and are therefore formally analogous to those that are generated by classical methods as discussed under 2(c)

Table I. Posterior probabilities of nonstationarity: data for figure 2 ($T = 50$)

	$\hat{\rho}$	$m(\hat{u})$	$m(y)$	$P_J(\rho \geqslant 1{\cdot}0)$	$P_F(\rho \geqslant 1{\cdot}0)$
Figure 2a					
(a) curves	0·804	33·62	78·49	0·5494	0·0209
(b) curves	0·990	58·99	2002·71	0·5250	0·3626

above. There is no 'disconcerting topology' here, simply genuine uncertainty about the generating mechanism, given the observed time-series. The J-posteriors manifest this uncertainty, the F-posteriors do not. Thus complaints about the disconcerting shape of confidence sets are as easily levelled against Bayes methods in practice as they are against classical theory. But this is a diversion from the real issue, which is the inherent uncertainty in time series estimation that results from the serial dependence on the data. Flat priors mask this uncertainty because they focus the posterior solely on the value of the fitted regression coefficient $\hat{\rho}$, just as if the data came from an independent sample with fixed regressors. In so doing they neglect the fact that we know *a priori* that the true value of ρ influences the autocorrelation structure of the time series and hence the anticipated amount of information that is carried by the data about ρ. By ignoring this information, flat priors are informative and, in consequence, they bias the posterior towards stationary, or more specifically, independent data alternatives.

The second set of curves, which are designated '(b)' in Figure 2, represent another typical outcome, in this case where the fitted regression coefficient $\hat{\rho}$ is close to unity. From Table I, we have $\hat{\rho} = 0{\cdot}99$. Both posteriors now attach an appreciable probability to the set $\{\rho \geqslant 1\}$ and thereby generally conform the data generating mechanism in both cases, although $P_J(\rho \geqslant 1)$ is still higher than $P_F(\rho \geqslant 1)$. The J-posterior is also unimodal, like the F-posterior, and the two densities are close in location as well as shape. Thus, for the sample outcomes given in (b) there is no great difference between the posteriors, and Bayesian methods as well as classical tests confirm the presence of a unit root.

5. As indicated above, the flat prior has a tendency to bias the posterior towards the i.i.d. alternative (i.e. $\rho = 0$ in (1)). By centering the posterior on $\hat{\rho}$ it will in any event inherit the downward bias of the regression estimator. But even when $\hat{\rho}$ is close to unity there may still be a non-negligible downward bias in the F-posteiror probabilities. For instance, in case (b) of Figure 2, in Table I we have a fitted coefficient $\hat{\rho} = 0{\cdot}99$ and yet $P_F(\rho \geqslant 1{\cdot}0) = 0{\cdot}3626$ which is substantially less than 50%.

The extent of the bias that is on average transmitted to the F-posterior can be measured by computing the expected posterior probability of the nonstationary set $\{\rho \geqslant 1\}$. This is easily done by simulation, and we found the following estimates of these expected probabilities for the case $T = 50$ from 20,000 replications:

$$E\{P_F(\rho \geqslant 1)\} = 0{\cdot}389,\ E\{P_J(\rho \geqslant 1)\} = 0{\cdot}625, \tag{12}$$

which confirm the downward bias of the F-posterior.

3.3. The AR(1) with Fitted Intercept and Trend

The methods of the previous section that employ ignorance priors may be used in much more complicated time series models. We shall illustrate the ideas first by extending the analysis to a model with a fitted intercept and trend, i.e.

$$y_t = \mu + \beta t + \rho y_{t-1} + u_t,\ u_t \equiv \text{i.i.d. } N(0, \sigma^2). \tag{13}$$

We choose this particular parameterization, rather than the one used by Sargan and Bhargava (1983) and by Schmidt and Phillips (1989), because it will facilitate comparisons with earlier work, especially when it is extended to accommodate transient dynamics.

Solving for y_t in (13), we have for $\rho \neq 1$ (the value at $\rho = 1$ may be calculated directly or

by means of l'Hopital's rule)

$$y_t = \sum_0^{t-1} \rho^t u_{t-t} + \mu(1-\rho^t)/(1-\rho) + \beta\{t/(1-\rho) - \rho(1-\rho^t)/(1-\rho)^2\} + \rho^t y_0,$$

and when $y_0 = 0$,

$$\begin{aligned} E(y_t^2) &= \sigma^2(1-\rho^{2t})/(1-\rho^2) + \mu^2\{(1-\rho^t)/(1-\rho)\}^2 + \beta^2\{t/(1-\rho) - \rho(1-\rho^t)/(1-\rho)^2\}^2 \\ &\quad + 2\mu\beta\{(1-\rho^t)/(1-\rho)\}\{t/(1-\rho) - \rho(1-\rho^t)/(1-\rho)^2\} \\ &= \sigma^2\alpha_{0t}(\rho) + \alpha_{1t}(\rho, \mu, \beta), \text{ say.} \end{aligned}$$

Summing over t we have

$$\sum_1^T E(y_{t-1}^2) = \sigma^2\alpha_0(\rho) + \alpha_1(\rho, \mu, \beta),$$

where

$$\alpha_0 = \alpha_0(\rho) = \sum_1^T \alpha_{0t-1} = T(1-\rho^2)^{-1} - (1-\rho^2)^{-2}(1-\rho^{2T}), \tag{14}$$

$$\begin{aligned} \alpha_1 = \alpha_1(\rho, \mu, \beta) &= \sum_1^T \alpha_{1t-1} \\ &= \sum_0^{T-1} [\mu(1-\rho)^{-1}(1-\rho^t) + \beta\{(1-\rho)^{-1}t - \rho(1-\rho)^{-2}(1-\rho^t)\}]^2. \end{aligned} \tag{15}$$

Again, these formulae yield the correct results by l'Hopital's rule when $\rho = 1$. The diagonal element corresponding to ρ of the information matrix for the model (13) is then

$$\sigma^{-2} \sum_1^T E(y_{t-1}^2) = \alpha_0(\rho) + \alpha_1(\rho, \mu, \beta)/\sigma^2.$$

The diagonal elements of the information matrix corresponding to μ, β and σ^2 are, respectively, $\sigma^{-2}T$, $\sigma^{-2}T(T+1)(2T+1)/6$ and $\sigma^{-2}2T$. Rather than work with the determinantal form of the Jeffreys prior (3), it is most convenient here to use the product of the diagonal elements of the information matrix. This leads to the following form of the ignorance prior for the model (13):

$$\pi(\rho, \sigma, \mu, \beta) \propto \sigma^{-3}\{\alpha_0(\rho) + \alpha_1(\rho, \mu, \beta)/\sigma^2\}^{1/2}. \tag{16}$$

The prior (16) is graphed in Figures 1(i) and (ii) for $\sigma = 1$ and for various values of μ and β; and the log density is graphed in Figures 1(iii) and (iv) for a wider range of ρ values. These graphs display the same characteristics as those of the earlier ignorance prior (9) for the simple AR(1). As μ and β depart from zero, the prior (16) obviously increases. However, as shown in Figures 1(iii) and (iv) the proportional increase in the prior is greater for $\rho < 1$ than it is for $\rho \geqslant 1$. Thus, we anticipate that the introduction of deterministic components in the model puts, relatively speaking, more additional weight on stationary ρ than it does on nonstationary ρ.

Let $\gamma' = (\mu, \beta)$, $\delta' = (\rho, \gamma')$ and use y_{-1}, X and Z to represent the observation matrices of (y_{t-1}), $(1, t)$ and $(y_{t-1}, 1, t)$, respectively. Under a Gaussian likelihood the joint posterior for (ρ, σ, γ) is

$$p(\rho, \sigma, \gamma \mid y) \propto \pi(\rho, \sigma, \gamma)\sigma^{-T} \exp\left\{-(1/2\sigma^2) \sum_1^T (y_t - \mu - \beta t - \rho y_{t-1})^2\right\}. \tag{17}$$

We decompose the exponent sum of squares as

$$\sum_1^T (y_t - \mu - \beta t - \rho y_{t-1})^2 = m(\hat{u}) + (\delta - \hat{\delta})'Z'Z(\delta - \hat{\delta})$$
$$= m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y) + (\gamma - \tilde{\gamma})'X'X(\gamma - \tilde{\gamma}), \tag{18}$$

where

$$m(\hat{u}) = \sum_1^T \hat{u}_t^2, \hat{u}_t = y_t - \hat{\mu} - \hat{\beta} t - \hat{\rho} y_{t-1}$$

are the OLS residuals and

$$m_X(y) = y_{-1} Q_X y_{-1}, Q_X = I - X(X'X)^{-1}X'$$
$$\tilde{\gamma} = \hat{\gamma} + (X'X)^{-1}X'y_{-1}(\hat{\rho} - \rho).$$

The component form (18) is especially useful in marginalizing the joint posterior (17). Although the prior $\pi(\cdot)$ is an awkward function of the parameters, the posterior (17) may be easily marginalized using the Laplace approximation for multivariate integrals. This approach was developed and used by Phillips (1983) in earlier related work on marginalizing exact multivariate densities. The reader is referred to that paper for a detailed discussion of the technique in this context. The method has subsequently received a good deal of attention in the Bayesian literature (see, especially, Tierney and Kadane, 1986; Tierney, Kass, and Kadane 1989). It provides a convenient and effective alternative to simulation-based numerical integration. In the present case we use the method to integrate out γ from (17) as follows, noting that the major contribution to the integral arises from a neighbourhood of $\gamma = \tilde{\gamma}$,

$$p(\rho, \sigma | y) \propto \int_{\mathbb{R}^2} \pi(\rho, \sigma, \gamma)\sigma^{-T} \exp\{-(1/2\sigma^2)[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]\}$$
$$\times \exp\{-(1/2\sigma^2)(\gamma - \tilde{\gamma})'X'X(\gamma - \tilde{\gamma})\}\, d\gamma$$
$$\sim (2\pi)|X'X|^{-1/2}\pi(\rho, \sigma, \tilde{\gamma})\sigma^{-T+2} \exp\{-(1/2\sigma^2)[m(\hat{u}) + (\rho - \hat{\rho})m_X(y)]\}. \tag{19}$$

Since the elements of $X'X$ are at least $O(T)$, the approximation (19) has a relative error of $O(T^{-1})$. For our purposes this will generally be quite adequate.

It remains to marginalize (19) with respect to σ. The derivations are given in Appendix B and lead the following marginal posterior for ρ:

$$p_J(\rho \mid y) \propto \alpha_0(\rho)^{1/2}\eta(\rho)^{-T/2}\Psi(T/2, (T+3)/2;\ (1/2\eta(\rho))[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]), \tag{20}$$

where $\eta(\rho) = \alpha_1(\rho, \tilde{\gamma}(\rho))$ and Ψ is a confluent hypergeometric function of the second kind (e.g. Erdélyi, 1953, p. 255).

This is a useful but complicated analytic formula for the posterior density. It may be simplified considerably when the order of magnitude of the final argument of the Ψ function is known. In the illustration we shall consider below, the true values of the coefficients in (13) are $\beta = 0$, $\mu \neq 0$ and $\rho_0 = 1$. The model then delivers a stochastic trend with drift and the quadratic form $m_X(y) = O_p(T^2)$. For a range of values of ρ we find that

$$(1/2\eta(\rho))[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)] \tag{21}$$

is very large relative to the other arguments. In this case, as shown in Appendix B, there is a very simple approximation to the posterior (20) given by the following

$$p_J(\rho \mid y) \propto \alpha_0(\rho)^{1/2}[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]^{-T/2}. \tag{22}$$

Although this approximation to (20) does not hold uniformly in ρ, computations comparing (20) and (22) show that (22) it is quite satisfactory for our present purposes.

We make the following observations.

1. Formula (22) is the direct analogue of our earlier formula (11) for the posterior density of ρ in the AR(1). All that differs is that the regression from which $\hat{\rho}$ and $\hat{u}$ arise now involves an intercept and trend as in (13) and the sample sum of squares $m(y)$ in (11) is replaced by the sum of squares, $m_X(y)$, of the detrended data. Given the correspondence between (22) and (11) it might be thought that (22) could be obtained much more simply through the direct use of the prior $\pi(\rho) \propto \alpha_0(\rho)^{1/2}/\sigma$, which employs independent uniform priors on μ and β, in place of (16). However, such a prior leads to a posterior for ρ in which the power of the quadratic form in square brackets in (22) is $-(T-2)/2$ not $-T/2$, and this posterior is improper. The posterior (22), on the other hand, is proper and, like (11), has Cauchy-type tails.
2. In view of this correspondence, the remarks we have already made in Section 3.2 regarding the properties of (11) also apply to (22). In particular, the posterior density (22) is asymmetric, it can be bimodal and the confidence sets that it generates display considerable uncertainty about the true coefficient ρ_0. In each of these respects if differs from the posterior density obtained from a flat prior. The latter, like (5), has the form $\pi(\rho, \sigma, \gamma) \propto 1/\sigma$ and we may therefore integrate out both γ and σ directly leading to the posterior density
$$p_F(\rho \mid y) \propto [m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]^{-(T-2)/2}. \tag{23}$$
This density, like (7), is symmetric about the regression estimate $\hat{\rho}$. As before, it inherits the bias of $\hat{\rho}$. But this bias is more severe in models with a fitted trend such as (13) than it is for the simple AR(1). We can therefore expect confidence sets that are based on (23) to exhibit a stronger downward bias than similar confidence sets from models with no fitted trend.

   Note that (23), unlike (22), involves a degrees of freedom adjustment in the exponent resulting from the marginalization with respect to γ and the use of a uniform prior. A common critique of the Jeffreys prior, originally expressed by Jeffreys (1961, p. 182) himself, is that it leads to no such degrees of freedom adjustment in multivariate contexts, at least without modification. Note, however, that were such an adjustment to be made to (22), the resulting posterior would be non integrable, as discussed in (i) above. In consequence, the degrees of freedom differential between (22) and (23) is quite justified in this context.
3. Figure 3 illustrates typical shapes for the posterior densities $p_J(\rho \mid y)$ and $p_F(\rho \mid y)$ for data generated from (13) with $\mu = 0{\cdot}025$, $\beta = 0{\cdot}0$, $\sigma^2 = 1$, $\rho = 1$ and $T = 50$. Two different data sets are used and the sample characteristics are given in Table II. The (a) curves in Figure 3 show a typical outcome where $\hat{\rho} = 0{\cdot}801$ is low. The J-posterior is bimodal and gives a posterior probability, $P_J(\rho \geqslant 1)$, to the nonstationary set of 7 per cent. The F-posterior is centered on $\hat{\rho}$ and gives only a 2 per cent probability to a stochastic nonstationary process. The (b) curves show a typical outcome where $\hat{\rho}(= 0{\cdot}974)$ is close to unity. In this case both posteriors give an appreciable probability to the presence of a stochastic trend, although $P_J(\rho \geqslant 1{\cdot}0)$ is substantially greater than $P_F(\rho \geqslant 1)$.
4. Expected posterior probabilities of the nonstationary set $\{\rho \geqslant 1\}$ were computed by simulation. From 20,000 replications using the model (13) with $\mu = 0{\cdot}025$, $\beta = 0{\cdot}0$, $\sigma^2 = 1$, $\rho = 1$ and $T = 50$, we found:
$$E\{P_F(\rho \geqslant 1)\} = 0{\cdot}0456,\ E\{P_J(\rho \geqslant 1)\} = 0{\cdot}2975. \tag{24}$$

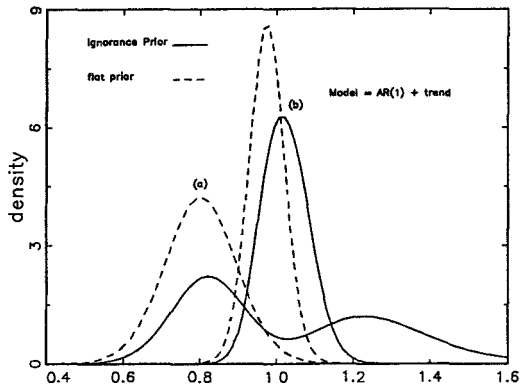

Figure 3. Posterior densities for ρ

Table II. Posterior probabilities of nonstationarity: data for figure 3

	Regression outcomes					Posterior probabilities	
	$\hat{\rho}$	$\hat{\mu}$	$\hat{\beta}$	$m(\hat{u})$	$m_X(y)$	$P_J(\rho \geqslant 1{\cdot}0)$	$P_F(\rho \geqslant 1{\cdot}0)$
(a) Curves	0·801	−0·228	−0·026	39·25	94·75	0·4658	0·0203
(b) Curves	0·974	−0·274	0·0175	45·79	463·60	0·6348	0·2996

Compared with the corresponding figures given in (12) for the simple AR(1) model, both expected posterior probabilities are smaller. But the J-posterior still gives an appreciable probability on average to $\{\rho \geqslant 1\}$, whereas the expected F-posterior probability is so small that inferences are certain to be biased away from finding evidence in support of a unit root. Indeed, in using the model (13) and flat priors for its coefficients, we must expect to find little evidence from the posterior distribution in support of a stochastic trend when such a trend is, in fact, present.

3.4. Models with Fitted Trends and Transient Dynamics

Empirical models typically employ a richer dynamic structure than (13). So, as a final illustration, we shall consider the following autoregressive model with fitted intercept and trend

$$y_t = \mu + \beta t + \Psi(L)y_t + u_t, u_t \equiv \text{i.i.d. } N(0, \sigma^2), \tag{25}$$

where $\psi(L) = \Sigma_1^k \psi_i L^i$. This formulation includes the empirical specifications used in Nelson and Plosser (1982), where $k \leqslant 6$, and the model used in the exercises conducted by DeJong and Whiteman (1989b), where $k = 3$.

It is convenient to employ the following alternative parameterization of (25)

$$y_t = \mu + \beta t + \rho y_{t-1} + \sum_1^{k-1} \varphi_i \, \Delta y_{t-i} + u_t, \tag{26}$$

where

$$\rho = \sum_1^k \psi_i, \tag{27}$$

is interpreted as the long-run autoregressive impact coefficient and where $\varphi_i = -\sum_{i+1}^{k} \psi_j$ $(i = 1, \ldots, k-1)$ are parameters of the transient dynamics. If $\psi(L) = 0$ has a unit root, then $\rho = 1$ and (26) is the parametric specification used by Nelson and Plosser in conducting classical augmented Dickey–Fuller tests for the presence of a unit root.

As an approximation to a Jeffreys prior for the parameters of (26), we shall use

$$\pi(\rho, \sigma, \mu, \beta, \varphi) \propto \sigma^{-k-2}\{\alpha_0(\rho) + \alpha_1(\rho, \mu, \beta)/\sigma^2\}^{1/2}, \tag{28}$$

where $\varphi' = (\varphi_1, \ldots, \varphi_{k-1})$. This may be interpreted as an approximation to the square root of the product of the diagonal elements of the information matrix for $\theta = (\rho, \sigma, \mu, \beta, \varphi)'$. The approximation is based on the value of this product when $\varphi = 0$. Moreover, when $k = 1$, (28) reduces to the earlier expression (16) for the ignorance prior in the model (13). However, since it fails to take into account the time series effects of the parameters φ and their impact on the information matrix, the prior (28) is not a true ignorance prior except when $\varphi = 0$. For values of φ very different from zero, we would expect this to lead to bias to the extent that (28) is based on generic prior information concerning a model in which $\varphi = 0$. Thus, like the flat prior for the coefficient ρ in model (1), the prior (28) will be an 'informative' prior in model (27) when the transient dynamics play a major role in explaining the data. An adequate methodology for dealing with this extra degree of complication is now under development and will be reported elsewhere.

Let $y(0) = (y_0, \ldots, y_{-k+1})$ be the vector of initial values for (25), let V be the matrix of observations of $(1, t, \Delta y_{t-1}, \ldots, \Delta y_{t-k+1})$ and let $\delta = (\mu, \beta, \varphi_1, \ldots, \varphi_{k-1})' = (\gamma', \varphi')'$ be the corresponding vector of parameters. Then the joint posterior density for (ρ, σ, δ) is:

$$p(\rho, \sigma, \delta \mid y, y(0)) \propto \pi(\rho, \sigma, \delta)\sigma^{-T} \exp\left\{-(1/2\sigma^2) \sum_{1}^{T} \left(y_t - \mu - \beta t - \rho y_{t-1} - \sum_{1}^{k-1} \varphi_i \Delta y_{t-i}\right)^2\right\}$$

$$= \pi(\rho, \sigma, \delta)\sigma^{-T}\exp\{-(1/2\sigma^2)[m(\hat{u}) + (\rho - \hat{\rho})^2 m_V(y)]\}\exp\{-(1/2\sigma^2)(\delta - \tilde{\delta})' V' V(\delta - \tilde{\delta})\}, \tag{29}$$

where

$$\tilde{\delta} = \hat{\delta} + (V'V)^{-1\prime} y_{-1}(\hat{\rho} - \rho),$$

$$m_V(y) = y'_{-1} Q_V y_{-1}, \; Q_V = I - V(V'V)^{-1\prime},$$

$$m(\hat{u}) = \sum_{1}^{T} \hat{u}_t^2$$

and $\hat{u}_t = y_t - \hat{\mu} - \hat{\beta} t - \hat{\rho} y_{t-1} - \sum_{1}^{k-1} \hat{\varphi}_i \Delta y_{t-i}$ are the OLS residuals.

We now marginalize (29) with respect to δ using the Laplace approximation described in the previous section and subsequently marginalize with respect to σ, leading to the following marginal posterior for ρ:

$$p_J(\rho \mid y) \propto \alpha_0(\rho)^{1/2} \eta(\rho)^{-T/2} \Psi(T/2, (T+3)/2; (1/2\eta(\rho))[m(\hat{u}) + (\rho - \hat{\rho})^2 m_V(y)]) \tag{30}$$

which may be approximated by

$$p_J(\rho \mid y) \propto \alpha_0(\rho)^{1/2} [m(\hat{u}) + (\rho - \hat{\rho})^2 m_V(y)]^{-T/2} \tag{31}$$

when the third argument of Ψ is large.

We note the following:

1. The marginal density (31) has the same form as our earlier formulae (22) and (11) for simpler models. It has the convenience of being applicable for an arbitrary choice of autoregressive order k in (27). Again, it is integrable and has Cauchy-type tails.

2. The posterior density for ρ corresponding to the flat prior $\pi(\rho, \delta, \sigma) \propto 1/\sigma$ is

$$p_F(\rho \mid y) \propto [m(\hat{u}) + (\rho - \hat{\rho})^2 m_V(y)]^{-(T-k-1)/2} \tag{32}$$

and this density has properties analogous to those of (23). Observe again the degrees of freedom differential in the exponents of (31) and (32), arising for the same reasons as those given earlier in Section 3.3.

3. Simulation exercises reported in Phillips (1990) show that the posterior densities (31) and (32) have similar characteristics to those described earlier for the model without transient dynamics. The main effect of the presence of the extra regressors that capture the transient dynamics is to depress the estimated regression coefficient $\hat{\rho}$, relative to a model without these extra regressors, and correspondingly to reduce the posterior probabilities of $\{\rho \geqslant 1\}$ computed from the F-posterior (32). The J-posterior (31) frequently manifests bimodality, as it does in the simpler model, and this partly compensates for the reduction in the size of $\hat{\rho}$ when computing posterior probabilities of $\{\rho \geqslant 1\}$.

3. The autoregressive formulation (25) is a popular agnostic model for accommodating transient dynamics. It is clearly of interest to study the effects of using this model for inference when the data follow an explicit ARMA structure. The analytic form of the Bayesian posteriors (31) and (32) makes it very convenient for us to do this. Suppose, for instance, that the errors on (13) follow an MA(1) leading to the revised model

$$y_t = \mu + \beta t + \rho y_{t-1} + u_t, u_t = e_t + \theta e_{t-1}, e_t \equiv \text{i.i.d. } N(0, \sigma^2) \tag{13}'$$

and we use this model to generate data for various values of θ, while the more convenient AR model (26) is used for inference. We use $\rho = 1{\cdot}0$, $\mu = 0{\cdot}025$, $\beta = 0{\cdot}0$ and $\theta \in \{-0{\cdot}8, 0{\cdot}8\}$ in (13)′ and $k = 3$ in (26) to illustrate the effects of this extension.

Table III provides simulation results for the expected posterior probabilities of $\{\rho \geqslant 1\}$ from 20,000 replications when $T = 100$ for different values of θ. In all cases the F-posterior probability leads to inferences that are biased away from models with stochastic trends. The expected J-posterior probability of $\{\rho \geqslant 1\}$ is more consonant with the true data-generating mechanism for each value of θ. But we notice that its value is sensitive to θ, especially as θ becomes large and negative. Indeed, for $\theta = -0{\cdot}8$ the posterior probability of $\rho \geqslant 1$ is on average unity. This outcome is the result of the bias, discussed earlier in connection with the prior (28), that results from the fact that (28) is no longer an ignorance prior when $\varphi \neq 0$. As θ in (13)′ approaches the value $-1{\cdot}0$, the true data-generating process when $\beta = 0{\cdot}0$ and

Table III. Expected posterior probabilities of ρ (model (13)′, $T = 100$)

θ	$E[P_F(\rho \geqslant 1)]$	$E[P_J(\rho \geqslant 1)]$
−0·8	0·000	0·999
−0·6	0·012	0·993
−0·4	0·033	0·914
−0·2	0·044	0·678
0·0	0·046	0·395
0·2	0·049	0·242
0·4	0·054	0·192
0·6	0·063	0·183
0·8	0·072	0·188

$\rho = 1 \cdot 0$ tends to

$$y_t = \mu t + e_t. \tag{13}''$$

In this case the prior (28), which is flat for φ, effectively downweights trend stationary alternatives such as (13)″ in favour of difference stationarity. A true ignorance prior would take into account that confidence sets for ρ are substantially different for MA coefficients θ around -1.00 compared with those around $\theta = 0 \cdot 0$. Indeed, in a classical setting with $\rho = 1 \cdot 0$ and $\theta = -1 \cdot 0$ the coefficients ρ and θ are strictly unidentified in an ARMA(1, 1).

4. EMPIRICAL APPLICATION TO THE NELSON–PLOSSER SERIES

We apply the methodology of the previous section to the historical time series studied by Nelson and Plosser (1982). For each of the 14 series we obtain the F-posterior and J-posterior for ρ from a fitted model of the form (27). Nelson and Plosser chose values of k in the range $1 \leqslant k \leqslant 6$, and DeJong and Whiteman (1989b) in their reconsideration of these data chose $k = 3$ for all series. Since our approach uses analytic methods rather than simulation it is easy and convenient to compute posteriors for an entire family of empirical model specifications. An investigator who wished to use our methods could indeed do so, and even incorporate a prior distribution on the lag length parameter in the analysis. However, we shall not go as far as this in our present illustration. Instead, we shall report results for both $k = 1$ and $k = 3$ to achieve some comparability with earlier work, and to illustrate the impact of different time series specifications on Bayesian inference.

Figures 4(i)–(xiv) give the posterior densities of ρ for the series. In each figure the two solid lines represent the J-posterior computed from ignorance priors using the AR(3) and AR(1) models, coded '(a)' and '(b)', respectively; the dashed line gives the F posterior computed for the AR(3) model—it may be regarded as a smooth and untruncated approximation to the posterior of the largest autoregressive root given by DeJong and Whiteman (for $k = 1$ the approaches would be equivalent apart from the truncation). Table IV reports the posterior probabilities of nonstationarity ($\rho \geqslant 1$) and near nonstationarity ($\rho \geqslant 0 \cdot 975$) for each series and for each fitted model.

The observed differences in the posterior distributions are major, especially between the use of the AR(1) and AR(3) models, showing that time series specifications have an important

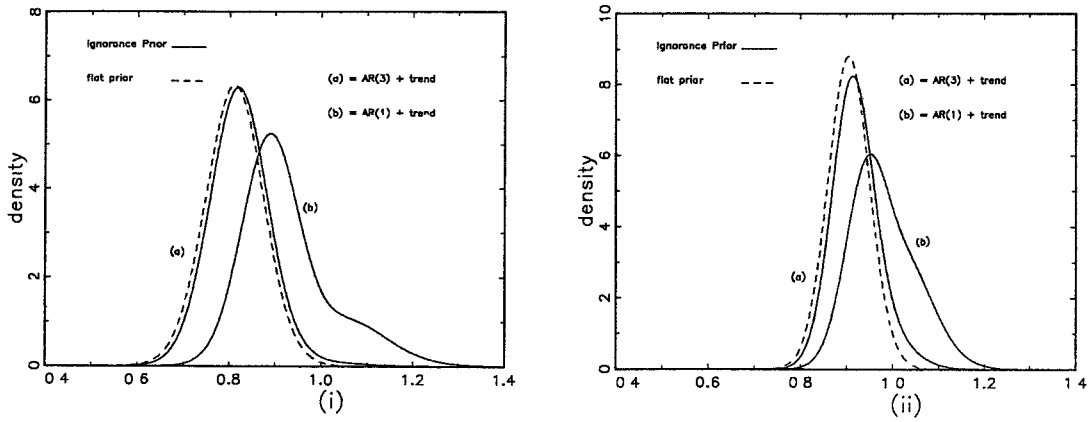

Figure 4. (i) Real GNP: 1909–1970, (ii) Nominal GNP: 1909–1970

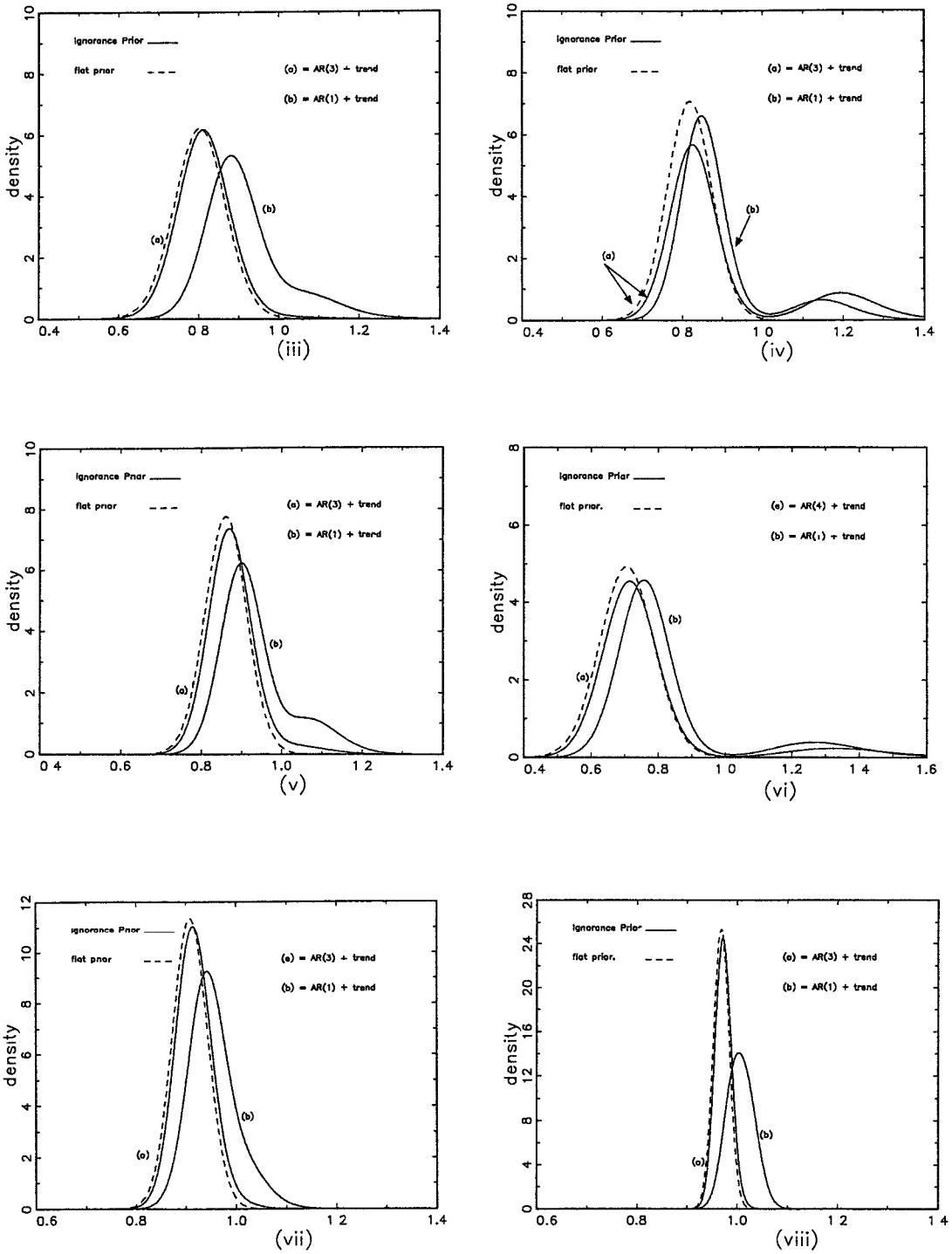

Figure 4. *continued.* (iii) Real per capita GNP: 1909–1970, (iv) Industrial production: 1860–1970, (v) Employment: 1890–1970, (vi) Unemployment rate: 1890–1970, (vii) GNP Deflator: 1889–1970, (viii) Consumer prices: 1860–1970.

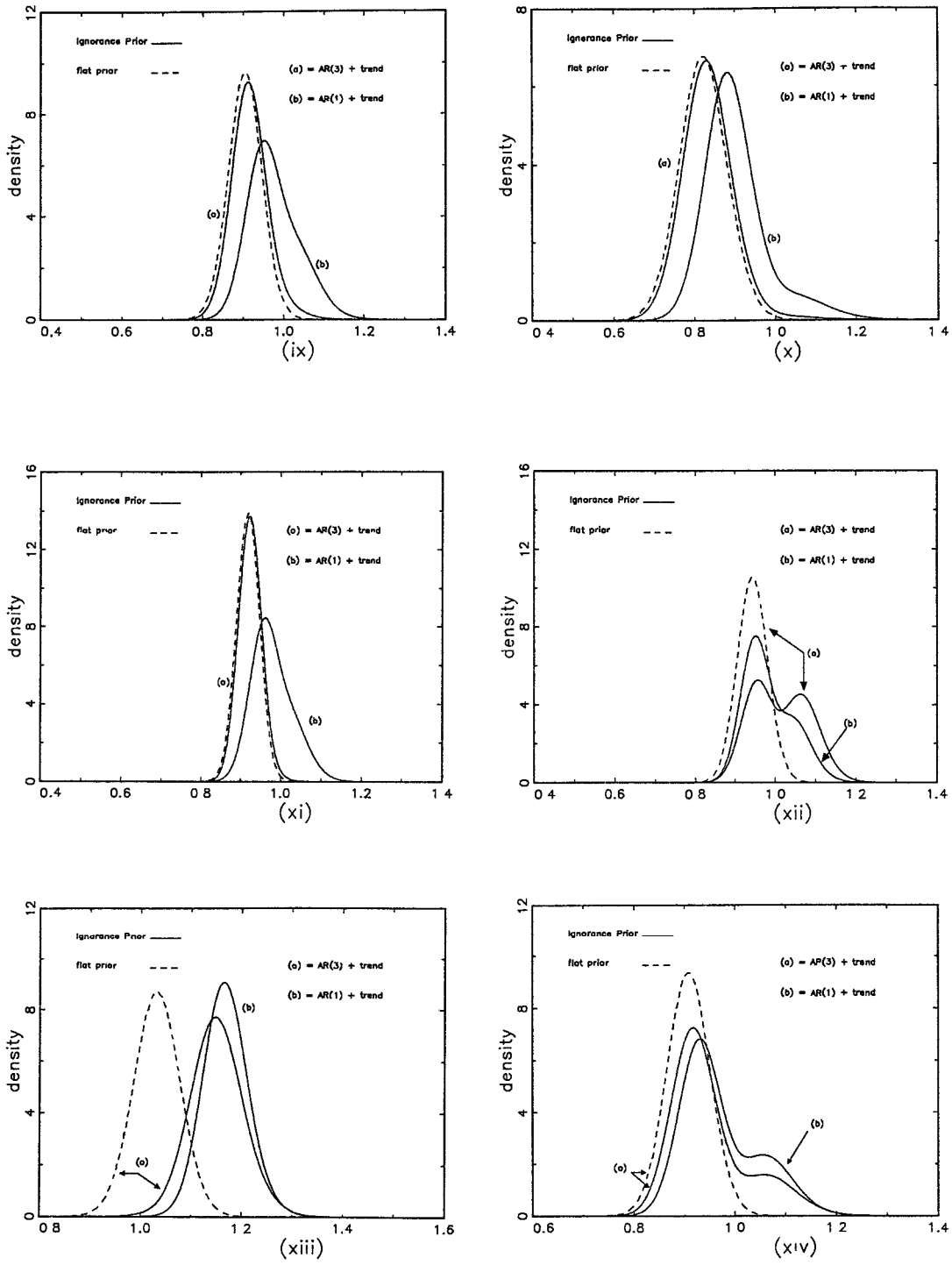

Figure 4. *continued.* (ix) Nominal wages: 1900–1970, (x) Real wages: 1900–1970, (xi) Money stock: 1889–1970, (xii) Velocity: 1869–1970, (xiii) Bond yields: 1900–1970, (xiv) Stock prices (SP500): 1871–1970.

Table IV. Posterior probabilities of stochastic nonstationarity

Series	AR(1) + trend				AR(3) + trend				
	$P_J(\rho \geq 1)$	$P_F(\rho \geq 1)$	$P_J(\rho \geq 0{\cdot}975)$	$P_F(\rho \geq 0{\cdot}975)$	$P_J(\rho \geq 1)$	$P_F(\rho \geq 1)$	$P_J(\rho \geq 0{\cdot}975)$	$P_F(\rho \geq 0{\cdot}975)$	$P_{DJW}(\Lambda \geq 0{\cdot}975)$†
Real GNP	0·193	0·023	0·242	0·054	0·012	0·002	0·019	0·005	0·003
Nominal GNP	0·361	0·092	0·485	0·203	0·074	0·021	0·141	0·063	0·020
Real per capita GNP	0·163	0·018	0·206	0·044	0·010	0·001	0·016	0·004	0·003
Industrial production	0·124	0·001	0·133	0·005	0·188	0·000	0·192	0·003	0·001
Employment	0·190	0·016	0·240	0·047	0·040	0·004	0·060	0·014	0·004
Unemployment*	0·126	0·000	0·129	0·001	0·086	0·000	0·087	0·000	0·002
GNP deflator	0·162	0·036	0·288	0·125	0·020	0·005	0·062	0·029	0·010
Consumer prices	0·601	0·272	0·880	0·713	0·176	0·082	0·652	0·528	0·196
Nominal wages	0·319	0·075	0·452	0·190	0·045	0·012	0·100	0·046	0·018
Real wages	0·103	0·011	0·140	0·031	0·014	0·001	0·021	0·005	0·003
Money stock	0·315	0·080	0·484	0·230	0·008	0·003	0·044	0·025	0·005
Velocity	0·353	0·051	0·483	0·168	0·537	0·073	0·642	0·204	0·592
Bond yields	0·999	0·968	0·999	0·992	0·996	0·764	0·998	0·892	0·617
Stock prices	0·301	0·028	0·385	0·092	0·215	0·017	0·278	0·059	0·040

* The penultimate four columns are based on an AR(4) + trend for this series, following Nelson and Plosser (1982).
† From Table 2 of DeJong and Whiteman (1989).

influence on posterior probabilities. For all series the J-posterior is located to the right of the F-posterior and attributes a greater probability to the nonstationary set $\{\rho \geqslant 1\}$. The J-posteriors are skewed to the right and for four series, notably industrial production (iv), the unemployment rate (vi), velocity (xii) and stock prices (xiv), they are bimodal. In the case of industrial production and the unemployment rate the bimodality arises in such a way that the main body of the distribution is located to the left of unity around the first mode and the density declines almost to zero between the modes. These two cases are very similar to the typical simulation outcomes given earlier in Figure 2. Like those cases, the bimodality here leads to disjoint shortest confidence sets and indicates substantial uncertainty about ρ. The bimodal posterior for velocity and stock prices takes a different form in that the density is substantial between the modes and confidence sets for ρ would not be disjoint. For these series there is less uncertainty about ρ and the posterior probability of nonstationarity is substantial in each case.

Table IV allows us to compare the posterior probabilities of nonstationarity for different model specifications and for flat prior and ignorance prior approaches. For the AR(1) + trend model with an ignorance prior, we have $P_J(\rho \geqslant 1) \geqslant 0{\cdot}30$ for seven series (nominal GNP, consumer prices, nominal wages, money stock, velocity, bond yields and stock prices) whereas for the same model with a flat prior, $P_F(\rho \geqslant 1) \geqslant 0{\cdot}30$ for only a single series (bond yields). For the AR(3) + trend model with our approximate ignorance prior (28), we have $P_J(\rho \geqslant 1) \geqslant 0{\cdot}15$ for five series (industrial production, consumer prices, velocity, bond yields and stock prices), whereas for the same model with a flat prior, we have $P_F(\rho \geqslant 1) \geqslant 0{\cdot}15$ again for only one series (bond yields).

It seems reasonable to conclude that, under conditions that approximate ignorance about ρ, there is substantially more evidence in support of stochastic trends than there is under an informative flat prior on ρ. Moreover, this conclusion appears robust to model specification. Note that for those series where the posterior probability of $\rho \geqslant 1$ may be considered small (less than 10 per cent, say) $P_J(\rho \geqslant 1)$ is still always greater than $P_F(\rho \geqslant 1)$ and often by a large multiple (usually 3–10 times greater). Thus, even in cases where inferences about nonstationarity are the same, the Jeffreys prior still has a large relative impact on the posterior probability.

Our empirical results under a flat prior on ρ are very similar to those reported in DeJong and Whiteman (1989b) for the dominant root, Λ, in the AR(3) characteristic equation. Their results were obtained by simulation-based numerical integration of the joint posterior and they base their inferences on the posterior probability of the near nonstationary set $\{\Lambda \geqslant 0{\cdot}975\}$. The final column of Table IV reports this probability as $P_{DJW}(\Lambda \geqslant 0{\cdot}975)$ and is taken from Table 2 of DeJong and Whiteman (1989b). Recall from our earlier discussion that DeJong and Whiteman employ a truncated flat prior on the autoregressive coefficients, but use Λ instead of ρ for their inferences. Since the flat prior is not invariant to parameter transformations, in contrast to the Jeffreys prior, their implied prior on Λ is not flat but is actually increasing in ρ. In spite of this, note that our $P_F(\rho \geqslant 0{\cdot}975) \geqslant P_{DJW}(\Lambda \geqslant 0{\cdot}975)$ for all of the series except velocity. DeJong and Whiteman infer from their results that evidence in support of a stochastic trend is present for only two series (velocity and bond yields) and they deem the evidence to be marginal in the case of a third series (consumer prices). An inspection of the penultimate column of Table IV, which reports our $P_F(\rho \geqslant 0{\cdot}975)$, shows that our methods support a similar inference. We differ by unequivocally including consumer prices, for which $P_F(\rho \geqslant 0{\cdot}975) = 0{\cdot}528$, in contrast to DeJong and Whiteman's $P_{DJW}(\rho \geqslant 0{\cdot}975) = 0{\cdot}196$. Only for these three series, *viz.* velocity, bond yields and consumer prices, are the posterior probabilities of $\{\rho \geqslant 0{\cdot}975\}$ and $\{\Lambda \geqslant 0{\cdot}975\}$ appreciable. For all other series the posterior

probability of a near nonstationary set is negligible: less than 6 per cent for $P_F(\rho \geqslant 0{\cdot}975)$ and less than 4 per cent for $P_{DJW}(\rho \geqslant 0{\cdot}975)$. We would not, of course, expect our flat prior results to be identical with those of DeJong and Whiteman because, as we have stated, they base their inferences on Λ not ρ and the implied prior for Λ, being a nonlinear function of the autoregressive coefficients, is not flat. However, this difference is simply one of parameterization. The starting point in the DeJong and Whiteman investigation is a flat prior on the autoregressive coefficients and in this respect is entirely analogous to our flat prior analysis.

Using flat priors, therefore, the evidence from the Nelson–Plosser time-series is that stochastic trends are unlikely for most of the series. Our results with ignorance priors show that these inferences based on flat priors are fragile for some of the series (especially stock prices) and they are always biased away from stochastic trend alternatives. The DeJong and Whiteman conclusions, we believe, should be interpreted with these qualifications in mind.

Although we have not discussed it above, DeJong and Whiteman also extract Bayesian posteriors for the trend coefficient, β, in (25) and, further, calculate the posterior for Λ under a prior for which $\beta = 0$. The latter exercise serves to highlight the role of the deterministic trend parameter in rendering stochastic trends 'unlikely' in their analysis. Our own analysis could be applied in a simple way and with no essential changes to perform exercises of this type. The remaining methodological development that would seem to be important for the use of our methods is the construction of a suitable Jeffreys prior which allows for the full transient dynamics in the general model (26). This is especially important because as k increases the transient dynamics soak up more of the observed variation in the series and inevitably reduce the role played by ρ or the largest autoregressive root Λ. Similar effects operate, of course, when there are more sophisticated deterministic trends in the model. A Jeffreys prior, which accommodates the information content of the moments of all of the regressors, should take these effects into account *a priori*. Work on this particular line of development is presently under way.

5. CONCLUSION

This paper set out to criticize recent Bayesian critiques of unit root econometrics. In so doing we have put forward an alternative Bayesian methodology based on the notion of ignorance priors, we have shown how to develop these priors for models where no stationarity assumption is made and we have shown how the methodology can be used in quite general autoregressive models with fitted trends. Our simulation exercises and our empirical application of these methods both indicate divergences that can be substantial from the results of a flat prior Bayesian analysis. This alone should be sufficient to alert us to the possibility of fragile inferences. But, as we have shown in addition, flat priors on the autoregressive coefficients are informative in time-series models, contrary to their apparent intent, and they typically downweight unit root and explosive alternatives in the posterior distribution. Moreover, as our illustrations demonstrate, Bayesian inferences are by no means robust to different time-series specifications and in some cases choice of lag length in an autoregression can have a major impact on inference. Finally, our simulation exercises and empirical results lead us to expect that an objective Bayesian analysis of stochastic trends will sometimes produce outcomes that are quite ambiguous due to a widely dispersed bimodality in the posterior distribution. In these cases, Bayesian methods reproduce in their own way a type of uncertainty that we normally associate with low discriminatory power in classical statistical

tests. Each of these factors should be borne in mind when interpreting Bayesian analyses of time series models.

In the light of these conclusions we submit that a Bayesian analysis of stochastic trends is by no means unequivocally superior to classical alternatives. Bayesian methods bring convenience and simplicity but also a host of issues that complicate inference in time series models and that go unmentioned in the Sims and Sims–Uhlig critiques. When these issues are ignored, as they most certainly are in the mechanical use of flat prior Bayesian analysis, the risk of misleadingly precise and biased inferences about stochastic trends is unacceptably large. Potential users of Bayesian methods need to be alerted to these shortcomings. In our view, one of the roles of scientific criticism is to do just this. To echo in the present context the sentiments that T. S. Eliot expressed about literary criticism in his Convocation Address to the University of Leeds, one would like to hope that one's

> *critical writings may be less fired by enthusiasm but informed by wider interest and, one hopes, by greater wisdom and humility* (p. 26).

In criticizing the critics of unit root econometrics this essay has attempted to put forward a wider and more objective perspective on Bayesian inference in time series models. We make no bones about the fact that we disagree with the deconstructionism of Sims (1988) and of Sims and Uhlig (1988/1991), we find their arguments about classical methods to be in error and their prescription of flat prior Bayesian methodology to be flawed. But we do see value in a Bayesian approach to inference that properly acknowledges the limitations of the approach. And we see no reason why empirical researchers should not judiciously pursue this approach as well as classical methods. If these perspectives on unit root econometrics are found by others to be of interest then this essay will have served its purpose.

APPENDIX A

For the Gaussian AR(1) model (1) with $\sigma^2 = 1$ and with a parameter sequence $\rho = \rho_0 + T^{-1}h$ adjacent to $\rho_0 = 1$ we have the log-likelihood ratio

$$\begin{aligned}\Lambda_T(h) &= \ln\{\text{pdf}(y;\rho)/\text{pdf}(y;\rho_0)\} \\ &= -(1/2)\sum_1^T (y_t - \rho y_{t-1})^2 + (1/2)\sum_1^T (y_t - \rho_0 y_{t-1})^2 \\ &= h\left(T^{-1}\sum_1^T y_{t-1}\varepsilon_t\right) - (1/2)h^2\left(T^{-2}\sum_1^T y_{t-1}^2\right).\end{aligned}$$

Under $\rho_0 = 1$ we have the following asymptotic behaviour established in Phillips (1987)

$$\Lambda_T(h) \xrightarrow{d} h\left(\int_0^1 W dW\right) - (1/2)h^2\left(\int_0^1 W^2\right) = \Lambda(h).$$

Under $\rho = \rho_0 + T^{-1}h$ we have the alternative limit

$$\Lambda_T(h) \xrightarrow{d} h\left(\int_0^1 J_h dW\right) - (1/2)h^2\left(\int_0^1 J_h^2\right),$$

from Phillips (1989), where $J_h(r) = \int_0^1 e^{(r-s)h} dW(s)$ is a diffusion process and $W(r)$ is

standard Brownian motion. Observe that under the local alternative sequence

$$T^{-2}\sum_{1}^{T} y_{t-1}^2 \xrightarrow[d]{} \int_0^1 J_h^2, \tag{A1}$$

a random limit which itself depends on h through the diffusion $J_h(r)$. In this sense the Fisher information is both random and variable (i.e. dependent on local departures) in the limit. The usual local asymptotic quadratic approximation does not apply. Because of this complication, the optimal asymptotic theory of inference of LeCam (1960, 1986) and Jeganathan (1980) is inapplicable in models with fitted unit roots. However, as shown in the authors (1988/1991) paper, these objections do not apply to models that are transformed to stationary form by differencing and cointegrating transformations.

APPENDIX B

We show how to marginalize the joint posterior (19). As in (19) we use the notation $\tilde{\gamma} = \tilde{\gamma}(\rho)$ and we write the factor involving the prior as

$$\pi(\rho, \sigma, \tilde{\gamma}) = \sigma^{-3}\{\alpha_0(\rho) + \alpha_1(\rho, \tilde{\gamma})/\sigma^2\}^{1/2} = \sigma^{-3}(\alpha_0 + \tilde{\alpha}_1/\sigma^2)^{1/2}.$$

The required marginal posterior may now be written in integral form as

$$p(\rho \mid y) \propto \int_0^\infty (\alpha_0 + \tilde{\alpha}_1/\sigma^2)^{1/2}\sigma^{-T-1}\exp\{-(1/2\sigma^2)[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]\}\, d\sigma.$$

Let $z = 1/\sigma^2$, $\eta = \tilde{\alpha}_1/\alpha_0$ and then

$$\begin{aligned} p(\rho \mid y) &\propto \alpha_0^{1/2}\int_0^\infty (1+\eta z)^{1/2} z^{(T/2)-1}\exp\{-(z/2)[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]\}\, dz \\ &= \alpha_0^{1/2}\eta^{-T/2}\int_0^\infty (1+v)^{1/2} v^{T/2-1}\exp\{-(v/2\eta)[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]\}\, dv \\ &= \alpha_0^{1/2}\eta^{-T/2}\Gamma(T/2)\Psi(T/2, (T+3)/2; (1/2\eta)[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]), \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function and Ψ is a confluent hypergeometric function of the second kind (see Erdélyi, 1953, p. 255). Taking out the constant of proportionality and noting that $\eta = \tilde{\alpha}_1/\alpha_0 = \eta(\rho)$ since $\alpha_0 = \alpha_0(\rho)$ and $\tilde{\alpha}_1 = \alpha_1(\rho, \tilde{\gamma}) = \alpha_1(\rho, \tilde{\gamma}(\rho))$ are functions of ρ, we obtain the following marginal posterior for ρ

$$p_J(\rho \mid y) \propto \alpha_0(\rho)^{1/2}\eta(\rho)^{-T/2}\Psi(T/2, (T+3)/2;\ (1/2\eta(\rho))[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]),$$

as given in equation (20).

When

$$(1/2\eta(\rho))[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]$$

is large relative to the other arguments of the Ψ function, the following approximation applies (see Erdélyi, *op. cit.*, p. 278):

$$\Psi(T/2, (T+3)/2;\ (1/2\eta))[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)] \sim \{(1/2\eta)[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]\}^{-T/2}.$$

With this approximation, we deduce a very simple approximation to the posterior, viz.

$$p_J(\rho \mid y) \propto \alpha_0(\rho)^{1/2}[m(\hat{u}) + (\rho - \hat{\rho})^2 m_X(y)]^{-T/2}.$$

ACKNOWLEDGEMENTS

All of the computations and graphics reported herein were carried out by the author using programs written in GAUSS on a ZEOS 386 20 mhz machine. Three referees gave helpful and informative reactions on the first version of this paper, which was originally circulated in July 1990 as a Cowles Foundation Discussion Paper, see Phillips (1990). Dale Poirier and Chris Sims commented extensively on the earlier version and their reactions are also appreciated. Thanks go to Charles Nelson for supplying the data used in Section 4, to Mico Loretan for computing advice, to Vassilis Hajivassiliou for the use of his graphics programs GPLOT and LJPLOT, to Glena Ames for outstanding wordprocessing and to NSF for research support under grant no. SES 8821180.

REFERENCES

Berenblut, I. I., and G. I. Webb (1973). 'A new test for autocorrelated errors in the linear regression models', *Journal of the Royal Statistical Society*, Series B, 33–50.

Box, G. E. P., and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.

Box, G. E. P., and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*, Addison Wesley, London.

Cooley, T. B., and S. F. LeRoy (1985). 'Atheoretical macroeconometrics: a critique', *Journal of Monetary Economics*, **16**, 283–308.

DeJong, D. N., and C. H. Whiteman (1989a). 'Trends and cycles as unobserved components in US real GNP: A Bayesian perspective', *Proceedings of the American Statistical Association*.

DeJong, D. N., and C. H. Whiteman (1989b). 'Trends and random walks in macroeconomic time series: a reconsideration based on the likelihood principle'. Working Paper No. 89-4, Department of Economics, University of Iowa. To appear in *Journal of Monetary Economics*.

DeJong, D. N., and C. H. Whiteman (1989c). 'The temporal stability of dividends and stock prices: Evidence from the likelihood function'. Working paper No. 89-3, Department of Economics, University of Iowa.

Dickey, D. A., and W. A. Fuller (1981). 'Likelihood ratio statistics for autoregressive time series with a unit root', *Econometrica*, **49**, 1057–1072.

Durlauf, S. N. (1989). 'Output persistence, economic structure and the choice of stabilization policy', *Brookings Paper on Economic Activity*, **2**, 69–116.

Durlauf, S. N. (1990). 'Locally interacting systems, coordination failure and the behavior of aggregate activity'. Working Paper, Stanford University.

Eliot, T. S. (1961). 'To criticize the critic', Convocation Lecture at the University of Leeds, 1961 in *To Criticize the Critic and Other Writings*, Faber, London, 1965.

Erdélyi, A. (1953). *Higher Transcendental Functions*, McGraw-Hill, New York.

Friedman, M. (1953). 'The methodology of positive economics', in *Essays in Positive Economics*, University of Chicago Press, Chicago.

Geweke, J. (1988). 'The secular and cyclical behavior of real GDP in nineteen OECD countries, 1957–1983', *Journal of Business and Economic Statistics*, **6**, 479–486.

Hall, R. E. (1978). 'Stochastic implication of the life cycle-permanent income hypothesis', *Journal of Political Economy*, **86**, 971–987.

Hartigan, J. A. (1964). 'Invariant prior distributions', *Annals of Mathematical Statistics*, **35**, 836–845.

Hartigan, J. A. (1965). 'The asymptotically unbiased prior distribution', *Annals of Mathematical Statistics*, **36**, 1137–1152.

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

Jeffreys, H. (1946). 'An invariant form for the prior probability in estimation problems', *Proceedings of the Royal Society of London*, Series A, 186, 453–461.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edn, Oxford University Press, London.

Jeganathan, P. (1980). 'An extension of a result of L. LeCam concerning asymptotic normality', *Sankhya Series A*, **42**, 146–160.

Leamer, E. E. (1983). 'Let's take the con out of econometrics', *American Economic Review*, **73**, 31–44.

Leamer, E. E. (1988). 'Things that bother me', *Economic Record*, **64**, 344–359.

LeCam, L. (1960). 'Locally asymptotically normal families of distributions', *University of California Publications in Statistics*, **3**, 37–98.
LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer, New York.
Lehmann, E. L. (1990). 'Comment', *Statistical Science*, **5**, 82–83.
Lindley, D. V. (1961). 'The use of prior probability distributions in statistical inference and decision', *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability*, **1**, 453–468.
Lindley, (1990). 'The present position in Bayesian statistics', *Statistical Science*, **5**, 44–89.
Malinvaud, E. (1980). *Statistical Methods of Econometrics*, 3rd edn, North Holland, Amsterdam.
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 (1989). 'Statistical inference in regressions with integrated processes: Part 2', *Econometric Theory*, **5**, 95–131.
Perks, F. J. A. (1947). 'Some observations on inverse probability including a new indifference rule', *Journal of the Institute of Actuaries*, **73**, 285–334.
Phillips, P. C. B. (1983). 'Marginal densities of instrumental variables estimators in the general single equation case', *Advances in Econometrics*, **2**, 1–24.
Phillips, P. C. B. (1987). 'Time series regression with a unit root', *Econometrica*, **55**, 277–301.
Phillips, P. C. B. (1988/1991). 'Optimal inference in cointegrated systems', Cowles Foundation Discussion Paper No. 866R, Yale University; and *Econometrica*, **59**, 283–306.
Phillips, P. C. B. (1988b). 'Reflections on econometric methodology', *Economic Record*, **64**, 344–359.
Phillips, P. C. B. (1989). 'Partially identified econometric models', *Econometric Theory*, **5**, 181–240.
Phillips, P. C. B. (1990). 'To criticize the critics: an objective Bayesian analysis of stochastic trends', Cowles Foundation Discussion Paper No. 950.
Runkle, D. E. (1987). 'Vector autoregressions and reality', *Journal of Business and Economic Statistics*, **5**, 437–442.
Sargan, J. D. (1979). 'The Durbin–Watson ratio of the Gaussian random walk', Working Paper, London School of Economics.
Sargen, J. D., and A. Bhargava (1983). 'Testing residuals from least squares regression for being generated by the Gaussian random walk', *Econometrica*, **51**, 153–174.
Schmidt, P., and P. C. B. Phillips (1989). 'Testing for a unit root in the presence of deterministic trends', Cowles Foundation Discussion Paper No. 933.
Schotman, P., and H. K. van Dijk (1991). 'A Bayesian analysis of the unit root in real exchange rates', *Journal of Econometrics*, (forthcoming).
Sims, C. A. (1980). 'Macroeconomics and reality', *Econometrica*, **48**, 1–48.
Sims, C. A. (1982). 'Policy analysis with econometric models', *Brookings Papers on Economic Activity*, 107–152.
Sims, C. A. (1988). 'Bayesian scepticism on unit root econometrics', *Journal of Economic Dynamics and Control*, **12**, 436–474.
Sims, C. A., J. H. Stock and M. W. Watson (1990). 'Inference in linear time series models with some unit roots', *Econometrica*, **58**, 113–144.
Sims, C. A., and H. Uhlig (1988/1991). 'Understanding unit rooters: a helicopter tour', Federal Reserve Bank of Minneapolis Institute for Empirical Macroeconomics, Discussion Paper No. 4. To appear in *Econometrica*, 1991.
Stock, J. H., and M. W. Watson (1988). 'Variable trends in economic time series', *Journal of Economic Perspectives*, **2**, 147–174.
Thornber, H. (1967). 'Finite sample Monte Carlo studies: an autoregressive illustration', *Journal of the American Statistical Association*, **62**, 801–818.
Tierney, L., and J. B. Kadane (1986). 'Accurate approximations for posterior moments and marginal densities', *Journal of the American Statistical Association*, **81**, 82–86.
Tierney, L., R. E. Kass, and J. B. Kadane (1989). 'Approximate marginal densities of nonlinear functions', *Biometrika*, **76**, 425–433.
Welch, B. L., and H. W. Peers (1963). 'On formulae for confidence points based on integrals of weighted likelihoods', *Journal of the Royal Statistical Society, Series B*, **25**, 318–329.
Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.
Zellner, A. (1988). 'Causality and causal laws in economics', *Journal of Econometrics*, **39**, 7–21.