COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale University
New Haven, Connecticut 06520

Cowles Foundation Discussion Paper No. 985

COMMENT ON "TO CRITICIZE THE CRITICS," BY PETER C.B. PHILLIPS

by

Christopher A. Sims

July 1991

Comment by Christopher A. Sims on
"To Criticize the Critics" by Peter C. B. Phillips
June 1991

In his paper "To Criticize the Critics" [1991b], Peter Phillips discusses Bayesian methodology for time series models. He criticizes earlier work by me ([1988], primarily) and by me and Harald Uhlig [1990] and argues for particular choices of standard prior distributions and particular methods for arriving at approximate characterizations of posterior distributions. Though I disagree with many of his suggestions, I am encouraged to see the expansion of discussion along this line.

The main point that Uhlig and I set out to make, however, was that careful consideration of the implications of the likelihood principle suggests that much of the recent work under the "unit root" label in the econometrics literature is being incorrectly interpreted in practice. We pointed out that time series models with possible unit roots are one of the few domains within which the implications of a likelihood-principle approach to inference are different, even in large samples, from those of a classical hypothesis-testing approach. Phillips addresses this part of our paper only indirectly. Without accepting our view that a likelihood-principle approach is strictly preferable to an hypothesis-testing approach, he endorses the usefulness of Bayesian methods, at least when they are by his standards properly applied. He makes no claim, directly or by the example of his work on Bayesian methods in this paper, that classical asymptotic theory for unit root models is useful from a Bayesian perspective.

Phillips's paper contains several indirect counterarguments to our claims for Bayesian or likelihood-based approaches. One is that Bayesian methods using what he regards as a well-founded "objective" prior distribution yield results for posterior probabilities that are closer to p-values developed from sampling distributions. While this of course does not address the questions we raise about the foundations of inference, it does suggest that these issues may be of less practical importance than we were claiming. Another is that our claim that classical methods in a sense ignore information is the opposite of the truth -- the Bayesian methods, by "conditioning on sample information" -- ignore information.

The remainder of this comment first reasserts the broader claims of the earlier papers about the clash between the likelihood principle and asymptotic sampling theory in these models, explaining why I find Phillips's counterarguments unconvincing. It then takes up the question of how to choose a good prior and how to report results in these models. Here I think that Phillips has in some ways improved on earlier suggestions of mine for standardized nonflat priors, but that his suggestions themselves have drawbacks. I suggest lines for further improvement.

I. Principles of Inference

Uhlig and I pointed out that in the model we considered there is in every sample a "prior" that would rationalize classical p-values as posterior probabilities. However, no prior can be chosen before observing the data that will rationalize p-values in all possible (or indeed in more than a measure-zero set of) samples. This means that treating p-values as probabilities is incoherent: if you are willing to quote me odds on bets concerning the true value of ρ (the coefficient in a univariate AR) based on p-values after you have seen the sample, and if you are also willing to quote me odds, before you have seen the sample, on bets concerning the true value of ρ and possible characteristics of the sample, then I can construct schemes in which you pay me money with probability one.

Of course we are not making bets when we report statistical results to a scientific audience, but I believe that this incoherence means that p-values are for these dynamic models a bad reporting device. The natural intuitive interpretation of them as probabilities is unsound. While it is true that the prior *pdf*'s that rationalize p-values in individual samples are convex and upward-sloping like the Jeffreys prior, Uhlig and I showed that they nonetheless differ quite sharply in different samples. That is, there is no one prior that rationalizes treating p-values as probabilities. As far as I can see, Phillips's paper makes no argument against this point.

Uhlig and I tried to provide some intuition for how it can be that, despite the downward bias in the OLS estimate $\hat{\rho}$ of ρ, the likelihood for ρ turns out to be symmetric about $\hat{\rho}$. We argued that the downward bias in $\hat{\rho}$ is balanced by an upward bias generated by the smaller variance of $\hat{\rho}$ when ρ is large. This balancing upward "bias" is not a classical bias in $\hat{\rho}$ for fixed ρ. Nonetheless it is easy to understand why, when $\hat{\rho}$ has smaller variance for larger ρ, this should make our beliefs about ρ given an observation on $\hat{\rho}$ lean toward higher values of ρ. We claimed that emphasizing the classical bias in $\hat{\rho}$ alone amounted to ignoring information about how the variance of $\hat{\rho}$ varies with ρ.

Phillips counters that it is Bayesian procedures that ignore information about variance, because they "condition" on the estimated variance of $\hat{\rho}$. I can see no argument for this proposition in what Phillips has written. A casual reader might accept the proposition based on a semantic confusion. Statistical procedures, whether classical or Bayesian, sometimes are said to "condition on" some subset of observed variables. Most commonly this involves using the conditional distribution of endogenous variables given exogenous variables, even where the distribution of the exogenous variables involves some of the parameters that are being estimated. Conditioning on variables in this sense does ignore possibly useful information. Bayesian or other likelihood-principle based inference "conditions on" the entire sample. That is, it aims at conditional probability statements given observations, so that probabilities of events that have not occurred are treated as irrelevant to assessing uncertainty about parameters given what has occurred. The fact that the likelihood principle implies that inference should be conditional on the entire sample certainly does not mean that it implies ignoring all the information in the sample. Two different meanings of "to condition on" are at play here.

Uhlig and I considered an artificial experiment in which values of an autoregressive coefficient ρ were drawn at random from a uniform distribution over a wide interval. Our point was to concentrate on a situation where everyone, frequentist or Bayesian, would agree on the appropriate formulas for calculating conditional probability distributions. I assume Phillips accepts Bayes' rule as a formula for calculating conditional distributions from a given joint distribution -- this formula is not what is at issue between Bayesians and frequentists. Our claim that everyone should agree on the symmetric shape of the distribution of ρ conditional on $\hat{\rho}$ in this experiment is just a matter of agreement by everyone on Bayes' rule. When there is a well-specified physical random mechanism generating ρ, use of it in the mathematical role of a "prior" is not a matter of subjective choice. It is only when the mechanism generating ρ is not observed repeatedly or otherwise is ill-defined that Bayesians can disagree about what prior to use and classical statisticians can object to treating ρ as a random variable. Thus there can be no good argument for using a Jeffreys prior or any other distribution in place of the actual random mechanism generating ρ in our example. Phillips's claim to the contrary must represent a misunderstanding of the artificial experiment we described, though on rereading our description it seems unambiguous to me.

II. Other Methodological Issues

Phillips also makes other general methodological points that are unrelated to our possible disagreements over the foundations of inference. This section takes these up briefly, more to indicate the nature of my reaction than to give detailed responses, since these issues are somewhat peripheral to the main dispute.

Phillips's paper quotes Stock, Watson and me [1990], labeling the quotation S_3, to the effect that any hypothesis that can be tested after a dynamic model has been "transformed" using unit root tests and the like can be tested with the same asymptotic distribution theory in the untransformed model. Phillips adds to the quote an interpolation of his own defining the word "transformed". The quoted statement is correct if the "transformed model" is interpreted as being one in which all restrictions corresponding to order T^{-1} components of the parameter space, including cointegration restrictions, are imposed. Phillips is pointing out that with his interpretation of "transformed", the statement is untrue, since certain of the order T^{-1} restrictions can be estimated and tested with convenient distribution theory if others of the order T^{-1} components are imposed first. He is correct, and by being imprecise about what "transformed" meant we may have given a mistaken impression in our statement.

The paper gives a brief litany of arguments against use of VAR's estimated in levels. I find none of these arguments convincing. Of course if we often knew that unit root hypotheses were exactly true, the argument for imposing them a priori would be strong. In fact, unit root and cointegration restrictions are estimated from the data. That is, we approach the data not knowing whether to impose them, and unit root tests and cointegration tests are part of a process by which the data determine the model we finally use. Asymptotic theory tells us that in large samples the unit root and cointegration hypotheses will be so sharply determined by the data that randomness in the estimation process generated from testing these hypotheses and imposing them as restrictions is negligible. However, as is evident from the shapes of the likelihood functions in applied work and from the conflicting and puzzling results often produced by the application of unit root tests, actual economic data seldom produce samples that are "large" in the required sense. Uncertainty about unit roots and cointegration is not trivially small relative to other sources of estimation error.

This is a difficulty one can confront explicitly by working with a model in levels, where it emerges in possibly inconvenient distribution theory for a classical approach and as the need for careful thought about priors and boundaries of parameter spaces in a Bayesian approach. Or one can proceed with the approach of estimation-by-hypothesis-testing, in which one arrives at a model with a simplified distribution theory by ignoring the randomness in the final model arising from the preliminary unit root testing and cointegrating vector estimation. Phillips's arguments for the latter course are convincing only if one believes that it often happens that the randomness being ignored is unimportant, as asymptotic theory promises it eventually will be. My view is that in practice it is seldom unimportant. What Phillips suggests amounts to a recommendation in favor of rose-colored glasses.

The paper accuses VAR's estimated in levels of being subject to "simultaneous equations bias". What is being called simultaneous equations bias here is not an inconsistency (as "simultaneous equations bias" is in standard usage). I regard this as a minor point, in large part addressed by the considerations in the preceding paragraph. Here I just warn the reader that there is a special usage of terminology here and suggest careful reading of Phillips's [1991a] paper for an understanding of what is at issue.

VAR's in levels are accused of generating "arbitrary" impulse response functions, with a citation of Cooley and Leroy [1985]. VAR's in standard form are reduced forms, and require formal or informal identifying assumptions to yield substantive conclusions. In my 1980 paper I summarized results with triangularly orthogonalized impulse response graphs. Such response graphs can be useful for any linear multivariate dynamic model, regardless of whether it is a VAR estimated in levels or a model in error-correction form estimated with unit root hypothesis tests. They are useful partly because they often provide, nearly or exactly, responses to underlying behavioral disturbances. I pointed out in the 1980 paper that some of the orthogonalized responses I displayed corresponded precisely to estimated responses to monetary policy disturbances under natural rate theories then popular. I also displayed hypothesis tests for Granger causality that connected directly to an identification of the model as a wage-price system of a then popular Keynesian type. There was nothing arbitrary about these tests and interpretations, and Cooley and Leroy did not present any arguments against them. They argue against a mechanical version of VAR modeling that may never have been implemented. Furthermore, there is by now a literature (Bernanke [1986], Blanchard [1989], Blanchard and Quah [1989], Blanchard and Watson [1986], and Sims [1986]) showing how formally to obtain structural interpretations of VAR impulse responses when simple triangular orthogonalization does not suffice.

VAR impulse responses are said to be "imprecise", with a citation of Runkle [1987]. I wrote a response to Runkle's article that appears in the same place, which the reader can consult. The gist of it is that impulse responses estimated from VAR's in levels are subject to considerable uncertainty, but that they nonetheless are estimated sharply enough to give us useful information. By imposing unit root and cointegration restrictions as if they were known a priori to be correct, one can greatly reduce the *apparent* imprecision of estimated impulse responses, but to take this as an argument in favor of estimation-by-hypothesis-testing again amounts to an argument for using such procedures as rose-colored glasses.

Unit root hypotheses are admitted by Phillips to be more precise than strictly justified by economic theory, but he claims that this situation is not different from what we see generally in econometrics, with sharp hypotheses being used, for convenience, as proxies for theories whose predictions are actually fuzzier. The issue here, as in all these cases, is how sharp the theory's prediction is relative to the most important alternatives. Considerable effort in the unit root literature has gone toward distinguishing a trend-stationary from a unit root model of real GNP, for example. The trend-stationary model is naturally interpreted as allocating business-cycle fluctuations, usually taken as lasting 2-5 years, to the stationary component. In quarterly data, roots with absolute values as high as .9 generate effects that have decayed to a tenth their original size within 5 years. In monthly data the corresponding root size is .96. In actual macroeconomic sample sizes, the power of unit root tests against alternatives that are this persistent is often very low. In effect, the sharp unit root hypothesis, if it is taken as correct when it is acceptable as a null hypothesis, is being allowed to proxy for a set of predictions so fuzzy that it heavily overlaps the most interesting alternatives.

III. Problems with the Jeffreys Prior

It is a small minority of Bayesian statisticians that accepts the notion that Jeffreys priors represent "ignorance" in any reasonable sense. Bayesians come in varieties. One variety is purely subjectivist, believing that prior distributions should always be selected to represent the individual investigator's knowledge and uncertainty. Another variety sees scientific data analysis as distinct from decision-making, so that it should use priors that are easily described, standardized across applications, and reflective of knowledge about parameters that is common across likely readers of the research report. Either of these varieties of Bayesian might find it convenient to compute a Jeffreys prior, but in each case the claim that a Jeffreys prior represents ignorance is only heuristic or suggestive. A subjectivist might be interested in a Jeffreys prior as a starting point when he thinks he knows little about the parameters in question. A likelihood-reporting Bayesian might be interested in a Jeffreys prior as possibly convenient, standardized, and representative of a common pattern of beliefs.

But from both of these Bayesian perspectives the Jeffreys priors that Phillips constructs for possibly non-stationary time series models are unattractive. They depend very strongly on sample size, putting increasingly high weight on explosive models as sample size increases. If the true model is not explosive, evidence against explosiveness actually accumulates very quickly as sample size increases, so that with any fixed set of prior beliefs posterior probability on the unstable region quickly becomes very small. Because the Jeffreys "priors" change as they do with sample size, however, the rate of decline of posterior probability on instability when the true model is stable is quite low, and indeed the posterior *pdf* always declines only as $|\rho|^{-2}$, so that its mean and variance are always undefined.

What is the argument for using "priors" that change with sample size? Phillips does not give one, and I cannot see how one could bec constructed. A subjectivist Bayesian believes priors should reflect actual a priori beliefs of the investigator. Conceivably a Jeffreys prior in some sample size could approximate someone's actual beliefs, but the subjectivist would see no reason to alter these beliefs for other possible sample sizes. From a likelihood-reporting perspective, the sample-size dependence of these priors is perverse. A reader interested in using reported results to reach conclusions using his own prior must first unravel the effects of the Jeffreys prior on the likelihood. The unraveling takes a different form with every sample size and initial condition. Furthermore, since no reader will have a prior that varies with sample size, it must be rare that the Jeffreys prior is close to a reader's beliefs.

The heavy weight placed by the prior on extremely explosive models is also unreasonable. As I will discuss below, there is a case for a prior that rises as ρ approaches one. There are even some kinds of data (e.g. price level data) where explosive models are plausible. Even in these cases, though, there are degrees of explosiveness so high that, if data suggested them as maximum likelihood estimates but with the likelihood fairly flat, nearly all researchers would act as if the truth were a less explosive model.

IV. The Flat Prior

For a subjectivist, the argument for a flat prior is only an approximate computational argument. Whatever the subjective prior, so long as it is characterized by a continuous *pdf*, under usual regularity conditions the posterior will come to resemble the likelihood function (the flat prior posterior *pdf*). This argument applies to Jeffreys priors in usual applications as well, since in the usual context Jeffreys priors do not depend on sample size. The Jeffreys priors for this time series application, however, depend so strongly on sample size that this justification for them fails -- posteriors from these Jeffreys priors have infinite variance in every sample size, whereas posteriors from any proper subjective prior with continuous and bounded *pdf* eventually have finite variance that shrinks to zero with sample size.

A scientific-reporting perspective suggests that the aim of statistical research reports is to summarize the likelihood. Since a flat-prior posterior is just the likelihood normalized to sum to one, it has a direct appeal from this perspective. Every reader can combine the likelihood with his own prior without having first to unravel the effects of some other prior.

The flat prior also has the advantage that it allows easy combination of results from several independent studies of the same parameter -- the posteriors, being likelihood functions, are just multiplied. Use of any other prior requires that the effects of the prior on the results be removed before results are combined. Otherwise there is double-counting of "prior" information.

Nonetheless there can be an argument for use of other priors in standard statistical reporting procedures. In time series models often there is no a priori reason to be sure that lag lengths are short, yet models that imply complicated effects at long lags are less plausible than models with shorter lags. The likelihood for a model parameterized to allow the possible long lags will often have a peak at an implausible value of the parameter vector, implying long and complicated lag structures, while still implying substantial probability on a region of more plausible models with shorter lags. While of course a complete description of the likelihood would bring out both aspects of the likelihood, usually it is more convenient simply to use a prior (as suggested some time ago by Leamer [1972] and Shiller [1973]) that damps the likelihood down in the region of large effects at long lags. Similarly, in contexts where a range of models has been contemplated, including some with complex structure included mainly as checks on simpler, more plausible models, it will often be useful to use a prior that damps down the likelihood for more complicated models. It is also possible for a "flat" prior, adopted thoughtlessly for a particular parameterization, to carry unwanted strong implications. Of course once recognized, such implications can in principle be discounted in interpreting results, but they may be quite inconvenient.

V. Time Aggregation

In applications of the simple univariate time series model under consideration here, it is often realistic to suppose that the same economic behavior can be studied using data of different frequencies. Thus stock price behavior can be examined with data by transaction, by minute, by hour, by day, by week, etc. If a univariate AR specification is exactly correct at a small time unit with a coefficient ρ, it remains correct at a unit S times as long, with coefficient ρ^S. If the prior *pdf* f applies to the coefficient $\phi=\rho^S$ in the regression estimable from sampled data, then the implied prior for the coefficient ρ with the small-time-unit data is

$$\rho^{S-1}f\left(\rho^S\right) . \tag{1}$$

Thus if we have a flat prior on the coefficient in an AR for annual data, our corresponding prior on the coefficient estimated from monthly data should be ρ^{11}, a convex upward-sloping function of ρ. A prior of $e^{-\phi}$ at the annual level, implying a prior mean of one on ϕ and substantial probability on explosive models, implies that the prior for monthly data is proportional to

$$\rho^{S-1}\exp\left(-\rho^S\right) , \tag{2}$$

which is plotted in Figure 1 along with the corresponding annual-data prior. Note that while this prior rises as ρ approaches one, it also drops rapidly for ρ above one. There is one prior that remains invariant to choice of time unit -- a flat prior on $\log(\rho)$ over $\rho>0$ (i.e., a prior *pdf* of $1/\rho$).

I believe that data collection and modeling strategy usually lead economists to use data they think corresponds to a "fine" time unit relative to the dynamics of the phenomena they are studying. Subjective priors therefore will naturally put more probability near $\rho=1$ than near $\rho=0$ or $\rho=\infty$. As part of reporting likelihood, therefore, it may be useful to report results from a prior shaped something like Figure 1, since something like that is more likely to be close to some readers' subjective priors than is a flat prior. However, despite their apparently very non-flat shape, such priors will not have much effect on the posterior *pdf* in most samples. Note that even the posteriors with Jeffreys priors that Phillips reports have little effect on posteriors except in the highly unstable region of the parameter space, and they involve factors of order ρ^{T-2} in that region where Figure 1's time-aggregation prior involves only a factor of ρ^{11} -- much smaller unless T is on the order of 13. Since they are likely to have little effect except in small samples, it may often be justifiable to use flat priors as a matter of convenience even where these time-aggregation priors would in principle better represent prior beliefs.

VI. Deterministic Trend Components

While Phillips does point out that the priors required to justify treating p-values as probabilities have the same general shape as Jeffreys priors, he does not note that when trend terms are introduced into the model, the Jeffreys priors retain their same general shape, while classical asymptotic distribution theory shifts drastically. When the μ and β in Phillips's model with trend are non-zero, classical asymptotic theory yields Gaussian distributions for the estimated ρ even in the presence of a unit root. In this case the usual situation of asymptotic equivalence of Bayesian posterior probabilities and classical one-tailed p-values prevails.

Paradoxically, we know that the bias in OLS estimates of ρ is much worse in this case than it is in the model without constant or trend (see Andrews [1991] for documentation of this point). Phillips points out that the Jeffreys prior, though still putting heavy weight on explosive models, does so to a lesser extent in this model than in the model without constant or trend. While Bayesian methods do not in general lead to unbiased estimators in the classical sense, where the bias is large and pervasive, as it is here, a Bayesian analyst should investigate it to be sure that it is reasonable, not an inadvertent implication of a conventional prior. That economic data contain strictly deterministic trends is even more implausible than that they evolve with exactly unit roots. When we construct a model like the one Phillips discusses, with

$$y(t) = \mu + \beta t + \rho y(t-1) + \varepsilon(t) \tag{3}$$

and consider testing for "stationarity about trend" versus "stochastic trend", we are actually letting both the unit root hypothesis and the deterministic trend hypothesis proxy for classes of nearby models. Since the two classes of models are difficult to distinguish in usual economic sample sizes, I would rather see work that used richer models, being more explicit about the classes of models being compared. Nonetheless the model (3) is used enough to deserve discussion.

Andrews [1991] has recently shown how to derive median-unbiased estimators for the parameters in (3). Part of his approach to the problem is to work with the alternative parameterization

$$y(t) = \bar{\mu} + \bar{\beta}t + (1-\rho L)^{-1}\varepsilon(t) \quad . \tag{4}$$

For $|\rho|<1$, (3) and (4) are alternative parameterizations of the same model. However a flat prior on $\bar{\mu}$, $\bar{\beta}$, ρ implies a prior with *pdf* proportional to $(1-\rho)^{-2}$ for μ, β, ρ. When we restrict the model to the $\bar{\beta}=0$ case, the prior *pdf* on μ, ρ corresponding to a flat prior on $\bar{\mu}$, ρ is proportional to $(1-\rho)^{-1}$. The reason for the strong non-flatness of these induced priors on (3) is that a flat prior on the parameters of (3) implicitly asserts that as ρ approaches one from below, the deterministic trend component of y becomes larger at such a rate that it continues to dominate the observable variation in y, while a flat prior on the parameters of (4) makes no such assertion. For the recent applications of this model in economics, where deterministic trend and ρ near one are being regarded as competing explanations of the same phenomenon, a flat prior on (4) seems to accord better with actual prior beliefs of most investigators.

For (3), regardless of whether μ or β is zero, the Jeffreys prior converges, as sample size goes to infinity, to proportionality to

$$\frac{1}{\sqrt{1-\rho^2}} \tag{5}$$

on the interval with $|\rho|<1$. This is a proper prior on that interval and converges to infinity at $\rho=1$ more slowly than either the prior on (3) that emerges from taking a flat prior on (4)'s parameters or the one that emerges from a flat prior on (4) with $\beta\equiv0$.

Note that equation (4) does not make sense, at least without further discussion, when $|\rho|\geq1$. By conditioning on y(0), however, we can construct a more robust version of it:

$$y(t) = \bar{\mu} + \bar{\beta}t + \left(y(0)-\bar{\mu}\right)\rho^{t} + \sum_{s=0}^{t-1} \varepsilon(t-s)\rho^{s} . \tag{6}$$

Equation (6) makes sense for all values of ρ, it continues to imply a version of (3), and now a flat prior on the parameters of (6) implies a prior *pdf* proportional to $|1-\rho|^{-2}$ on the parameters of (3). The corresponding result for the $\beta \equiv 0$ case is $|1-\rho|^{-1}$.

While these extensions of the prior derived from a flat prior on (4) to the non-stationary case are like the Jeffreys prior in that they increase toward infinity as $\rho \to 1$ from below, they differ drastically in the nonstationary region. There they decrease rapidly as ρ moves above one, while the Jeffreys priors in finite samples increase sharply as ρ moves above one and have no limiting form in this region as sample size approaches infinity. It is also worth noting, though not important in most economic applications, that these priors differ from the limiting Jeffreys prior in not having a second singularity at $\rho=-1$.

Finally, $|1-\rho|^{-j}$ has a non-integrable singularity at $\rho=1$, both for j=1 and j=2. Since the likelihood conditional on y(0) is non-zero at $\rho=1$, this non-integrability persists into the posterior.

There are other ways of coming up with candidates for standardized priors. If one is willing to assume stationarity of the stochastic component of y, it is attractive to use the unconditional likelihood, thereby adding a factor of $\sqrt{(1-\rho^2)}$ to the likelihood. Since the flat prior on the parameters of (6) leads to a nonintegrable posterior for the parameters of (3), it is natural to consider using loose but proper priors on the parameters of (6) -- say independent normal, with the mean of $\bar{\mu}$ centered at y(0).

VII. Practical Conclusions

In the homogeneous model, with no constant or trend term, a flat prior seems sensible. The Jeffreys prior generated from finite-sample distributions depends on sample size in a paradoxical way and puts unreasonable high weight on explosive models. The limiting Jeffreys prior for $T\to\infty$ only exists in the stable region. Time aggregation considerations suggest that within the stable region a prior smoothly favoring coefficients closer to one is reasonable, but quantitatively shapes for such priors based on time aggregation appear to deliver results close to those for a flat prior in most samples.

In models with constant or constant and trend, (where the classical distribution theory is asymptotically normal and thus suggests no strong bias in Bayesian results with flat priors) the flat prior is less attractive. It implies credibility for stationary models in which the deviation of y(0) from the deterministic component of the model is very large relative to the model's steady state. Apparent "upward trends" can then be explained as transient convergence to a deterministic path that lies on one side of the data for much of the sample. Least squares estimates of (3) typically imply this kind of fitted model. Figure 2 shows results from this type of model for quarterly U.S. postwar GNP data. The estimated model has an autoregressive coefficient of .96, nearly two standard errors from one using the conventional estimate of standard error. But the model implies a large transient that kept the data below the trend line from 1948 through the early 60's.

It makes sense to deviate from a flat prior on the parameters of (3) to reflect skepticism about models displaying such large transients. However, there is no single natural or limiting prior that takes care of this problem. Figure 3 displays posteriors for ρ under four priors, labeled flat, Jeffreys, loose, and tight. The flat prior is flat on the parameters of (3) and is, naturally, centered on the OLS estimate of .96. Results for the monthly data prior of Figure 1 are not displayed because, over the range of ρ's used in Figure 3, the monthly data Figure 1 prior is effectively flat -- the posterior *pdf* is almost indistinguishable from the plot of the likelihood. The Jeffreys prior is the limiting form of the Jeffreys prior as $T\to\infty$, which exists only on the stable interval for ρ. Though it is not shown on the graph, we know that its posterior *pdf* would rise to infinity if we extended it toward 1. Note that in the relevant range it is a modest alteration of the original flat prior posterior.

The loose and tight posteriors are generated from priors that are flat in ρ and independent normal in $\bar{\mu}$ and $\bar{\beta}$ from (6). Both center $\bar{\mu}$ at y(0), with the tight prior giving $\bar{\mu}$ a prior standard deviation of .2 and $\bar{\beta}$ a prior standard error of .004. The loose prior is approximately what is obtained as these prior standard errors approach infinity. Clearly results for these priors differ from flat-prior results much more sharply than do those for the Jeffreys prior, and they differ among themselves according to their tightness. Though they are not displayed here, further interesting variations in the posterior can be obtained by changing the relative prior standard deviations on $\bar{\mu}$ and $\bar{\beta}$.

My own conclusion from these results is that apparent strong significance for constant and trend terms in autoregressive models should be regarded skeptically. Often their acceptance implies a model with a large transient component. When such large transients are inherently implausible, results from models excluding the deterministic components (while perhaps adding more parameters to the stochastic dynamics) are likely to give more reasonable results. Where large transients have some plausibility, there is no way to avoid thinking about what kind of transients, and how large, are plausible, and evaluating the results from flat priors in this light. Doing this systematically will bring us back to a proper prior on $\bar{\mu}$ and $\bar{\beta}$ in (6). In the Figure 2 case, we need to ask whether a "postwar effect" pushing the economy 20% below its deterministic growth path and persisting over 15 years is reasonable for the U.S. To me, it seems somewhat implausible. But similar results for the Japanese or German economies would be more believable. (Of course this all depends on what the estimated deterministic growth path might represent, discussion of which would take us too far afield.)

VIII. Overall Conclusions

Phillips's paper seems to me a truly seminal contribution. Despite (or because of?) my many disagreements with it, it has made me think further about inference in these models, and I am sure it has done and will do the same for other readers. Though I had earlier [1988] suggested candidate standard non-flat priors (of a "spike and slab" type) favoring a unit root and pointed out their dependence on the time unit, smooth priors favoring unit roots are more reasonable (if somewhat less tractable). The argument that flat priors are unreasonable when the model contains constant and trend is stronger than I had realized (though it is not an argument that Phillips makes directly). Figuring out how the insights being developed in the current debate about Bayesian methods in these simple models can be extended to practice in multivariate, multilag models is a major challenge. There are more, and more interesting, open research topics in this area than was apparent before Phillips stimulated this discussion.

References

Andrews, Donald W.K. [1991]. "Exactly Unbiased Estimation of First Order Autoregressive Unit Root Models," Cowles Foundation Discussion Paper 975, April.

Bernanke, B. [1986]. "Alternative Explanations of the Money-Income Correlation", in *Carnegie-Rochester Policy Series on Public Policy*, Amsterdam: North-Holland.

Blanchard, Olivier [1989]. "A Traditional Interpretation of Macroeconomic Fluctuations," *American Economic Review, 79*, 1146-1164.

Blanchard, Olivier and Danny Quah [1989]. "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review, 79*, 655-73.

Blanchard, Olivier and Mark Watson [1986]. "Are All Business Cycles Alike?" in *The American Business Cycle*, R.J. Gordon, ed. (Chicago: U. of Chicago Press).

Cooley, Thomas B. and Stephen F. Leroy [1985]. "Atheoretical Macroeconometrics: A Critique," *Journal of Monetary Economics 16*, 283-308.

Leamer, Edward E. [1972]. "A Class of Informative Priors and Distributed Lag Analysis," *Econometrica*, *40*, p.1059.

Phillips, Peter C. B. [1991a]. "Optimal Inference in Cointegrated Systems," *Econometrica 59*, 283-306.

Phillips, Peter C. B. [1991b]. "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends," to appear, *Journal of Applied Econometrics*.

Runkle, David E. [1987]. "Vector Autoregressions and Reality," *Journal of Business and Economic Statistics 5*, 437-442.

Shiller Robert J. [1973]. "A Distributed Lag Estimator Derived From Smoothness Priors," *Econometrica*, *41*, p.775.

Sims, Christopher A. [1980]. "Macroeconomics and Reality," *Econometrica*, *48*, 1-49.

____________________ [1986]. "Are Forecasting Models Usable for Policy Analysis," *Quarterly Review* of the Federal Reserve Bank of Minneapolis, *10*(1), 2-16.

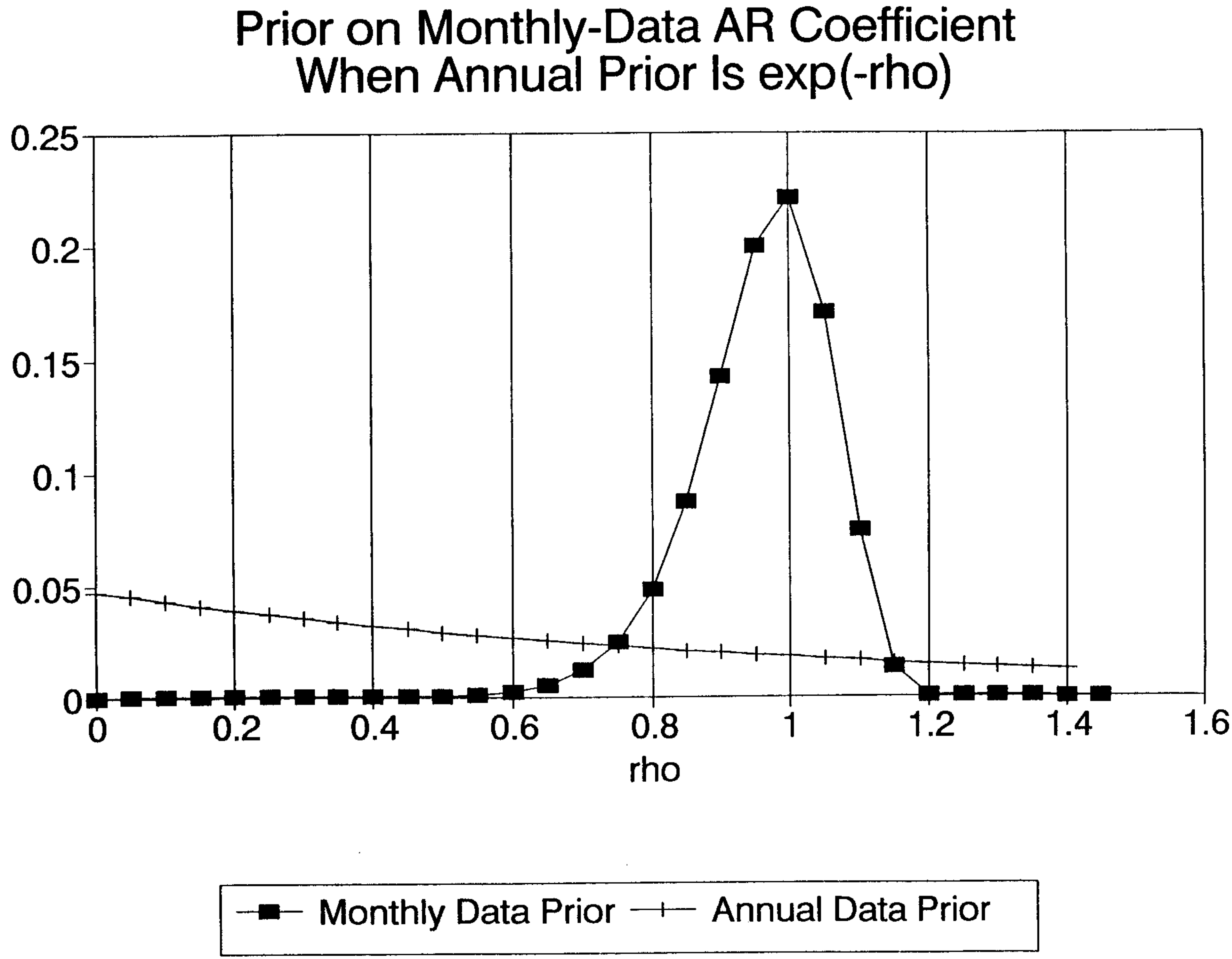
____________________ [1988]. "Bayesian Skepticism on Unit Root Econometrics," *Journal of Economic Dynamics and Control* **12**, 463-474.

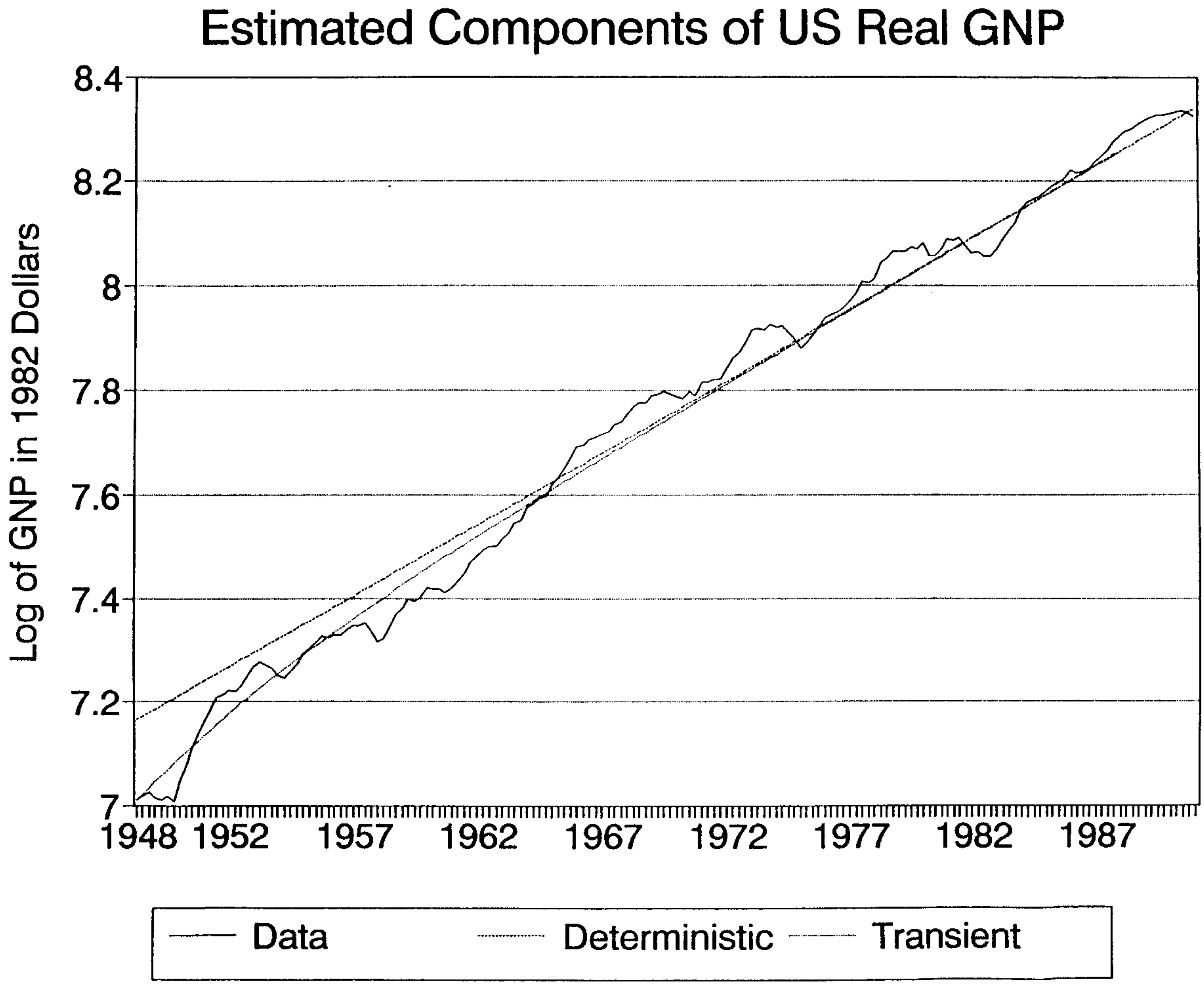
Sims, Christopher A. and Harald Uhlig [1991]. "Understanding Unit Rooters: A Helicopter Tour," to appear, *Econometrica*.

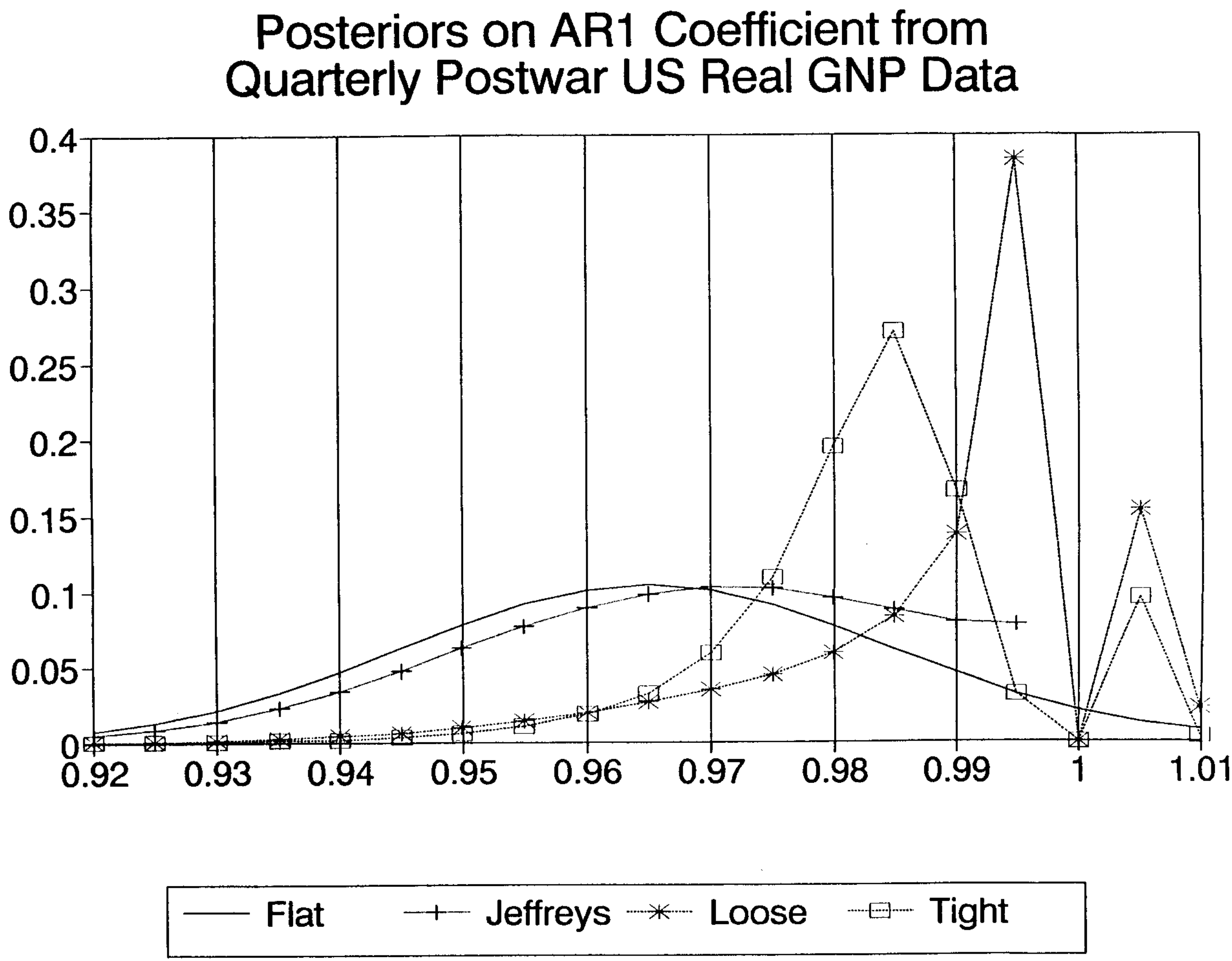
Sims, Christopher A., James H. Stock and Mark W. Watson [1990]. "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica 58*, 113-144.

Note to Figure 3

The posteriors were calculated by a naive numerical integration, except that the parameters were first transformed to correspond to an orthogonalization of the data matrix, The orthogonalizing transform was chosen to be an upper triangular matrix and the lagged dependent variable was put at the bottom of the parameter vector, so that it was unaffected by the transformation. The three-dimensional posterior *pdf* was then calculated at a grid of approximately 20 points in each dimension (about 8000 points in all) centered at the OLS estimates and covering about 3 standard errors on either side of the estimate. A similar grid for the untransformed parameter vector gave extremely inaccurate results for the marginal posterior for ρ, because likelihood is concentrated near a lower-dimensional submanifold of the original parameter space, meaning only a few of the grid points in that space have high likelihood. This problem is avoided by the preliminary orthogonalizing transformation.


Prior on Monthly-Data AR Coefficient
When Annual Prior Is exp(-rho)
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
rho
Monthly Data Prior
Annual Data Prior


Estimated Components of US Real GNP
Log of GNP in 1982 Dollars
8.4
8.2
8
7.8
7.6
7.4
7.2
7
1948
1952
1957
1962
1967
1972
1977
1982
1987
Data
Deterministic
Transient


Posteriors on AR1 Coefficient from
Quarterly Postwar US Real GNP Data
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.92
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1
1.01
Flat
Jeffreys
Loose
Tight