Econometrica, Vol. 67, No. 5 (September, 1999), 1113–1155

ERROR BANDS FOR IMPULSE RESPONSES

BY CHRISTOPHER A. SIMS AND TAO ZHA[1]

We show how correctly to extend known methods for generating error bands in reduced form VAR's to overidentified models. We argue that the conventional pointwise bands common in the literature should be supplemented with measures of shape uncertainty, and we show how to generate such measures. We focus on bands that characterize the shape of the likelihood. Such bands are not classical confidence regions. We explain that classical confidence regions mix information about parameter location with information about model fit, and hence can be misleading as summaries of the implications of the data for the location of parameters. Because classical confidence regions also present conceptual and computational problems in multivariate time series models, we suggest that likelihood-based bands, rather than approximate confidence bands based on asymptotic theory, be standard in reporting results for this type of model.

KEYWORDS: Vector autoregression, confidence region, Bayesian methods, impulse responses.

1. INTRODUCTION

IN INTERPRETING DYNAMIC MULTIVARIATE LINEAR MODELS, impulse response functions are of central interest. Presenting measures of the statistical reliability of estimated impulse responses is therefore important. We will discuss extensions of existing methods for constructing error bands that address some important practical issues.

- It has been conventional in the applied literature to construct a $1-\alpha$ probability interval separately at each point of a response horizon, then to plot the response itself with the upper and lower limits of the probability intervals as three lines. The resulting band is not generally a region that contains the true impulse response with probability $1-\alpha$ and does not directly give much information about the forms of deviation from the point estimate of the response function that are most likely. In Section 6 we suggest a way to provide such information.
- While there is a widely used, correct algorithm[2] for generating error bands for impulse responses in reduced form VAR models, it is not easy to see how to extend it to overidentified structural VAR's, and some mistaken attempts at extension have appeared in the literature. In Section 8 we show how correctly to make this extension and how to use numerical methods to implement it.

[1] Support for this research was provided in part by National Science Foundation Grant Number SES91-22355. The views expressed in this paper do not necessarily represent those of the Federal Reserve Bank of Atlanta or the Federal Reserve System.

[2] Distributed in the RATS manual since the earliest versions of that program.

But before we present the details of our proposed methods for dealing with these issues, we need to take up some conceptual issues. The error bands we discuss are meant to characterize the shape of the likelihood function, or of the likelihood function multiplied by the type of reference prior widely used in reporting results to a scientific audience. The importance of characterizing the shape of the likelihood, because of the likelihood's central role in any decision-making use of a statistical model, should not be controversial. The main critiques of likelihood-based inference (e.g. Loève (1988)) argue against the likelihood *principle*, the claim that *only* the likelihood need be reported, not against the importance of reporting the shape of the likelihood. Computing and reporting these likelihood-characterizing intervals is a substantial challenge in itself, and deserves more attention in applied work.

There are valid arguments for reporting more than the likelihood, and classical confidence intervals can be thought of as constructed from the likelihood plus additional information about overall model fit that is needed for model criticism. However, confidence intervals mix likelihood information and information about model fit in a confusing way: narrow classical confidence bands can be indicators either of precise sample information about the location of parameters or of strong sample information that the model is invalid. It would be better to keep the two types of information separate.

Section 2 lays out in more detail the general argument for reporting separately likelihood shape and fit measures, rather than reporting confidence intervals.

Section 3 defines ‘‘impulse responses’’ and the class of models we are considering.

Section 4 discusses the problems raised by the fact that we are forming measures of precision of our inference about vector-valued functions of a vector of parameters, with the dimensionality of both vectors high. For likelihood-describing error bands, these problems are just the need to make some choices about practical details of implementation. Attempts to instead form confidence regions run into serious conceptual problems in this situation.

It is not possible, even in principle, for the time series model in which we are interested, to construct classical confidence intervals for impulse responses with exact small-sample justification. There are a variety of approaches to obtaining approximate confidence intervals that grow more accurate in some sense as sample size increases. Many of these approaches that have been used in practice have the same degree of justification in first-order classical asymptotic theory as the practice of using Bayesian posterior probability intervals as if they were confidence intervals. There are particular bootstrap methods (little-used in applied time series research) that have improved second-order asymptotic properties in theory, though not in practice according to Monte Carlo studies. Section 5 discusses these methods and the problems in implementing them.

Examples in Sections 7 and 8B illustrate the performance of our suggested methods and provide some comparisons with other approaches.

2. LIKELIHOOD SHAPE VS. COVERAGE PROBABILITY

Formal decision theory leads, under some regularity conditions, to the conclusion that rational decision rules have the form of Bayesian rules.[3] Following Hildreth (1963), we might think of scientific reporting as the problem of conveying results of analysis to an audience that may have diverse prior beliefs about parameters and diverse loss functions, but accept a common statistical model. Reporting is not decision-making, and therefore makes no use of subjective prior beliefs. Instead it relies on the *likelihood principle*:[4] all evidence in a sample about the parameters of the model is contained in the likelihood, the normalized p.d.f. for the sample with the data held fixed and the parameters allowed to vary. Scientific reporting is then just the problem of conveying the shape of the likelihood to potential users of the analysis.

It does not seem reasonable, though, to suppose that very often a wide audience accepts a statistical model as certainly true, rather than as an interesting hypothesis or plausible approximation. Readers who are interested in the model and its parameters, but might want also to consider other models, will need more than the likelihood function. To compare the fit of one parametric model to another for the same data, one needs to construct the likelihood function for the grand model made up of the two, with a discrete parameter that chooses between the two models. To construct this overall likelihood, we need to use both the separate model likelihoods and their relative levels. If a proper[5] Bayesian reference prior is being used, this can be done by presenting the posterior p.d.f.'s for the two separate models, which will usually be close to their likelihoods in shape, and the posterior probabilities of the two models. Geweke (1995), for example, suggests reporting such information. But no use of a prior is necessary to report this information. The same information is contained in the two separate likelihoods themselves, together with any statistic that determines their relative heights—e.g. the ratio of p.d.f. values at some particular point in the sample space. Of course if the set of models to be considered is not known in advance, then the levels of the p.d.f. values are needed for all the models reported.

The likelihood function contains the sample information about parameters of the model, while the additional information about level of the p.d.f. contains information about fit of this model relative to possible competitor models. A classical confidence interval, together with its coverage probability, does not depend only on the likelihood. Therefore, in decision-making with a trusted model, use of a classical confidence interval and coverage probability must in general lead to suboptimal decisions. However, in scientific reporting, as we have just argued, it may be important to convey information beyond that in the likelihood. Therefore it cannot be argued that with the likelihood available,

[3] Such a result is called a Complete Class Theorem. See Ferguson (1967).

[4] See Berger and Wolpert (1988).

[5] I.e. a probability distribution, rather than a "density function" that does not integrate to one.

TABLE I

EMPTY OR TRIVIAL CONFIDENCE SETS

	A		B				
X	$P(X\|A)$	LR	$P(X\|B)$	LR	LR 100% conf. set	Simple 99% conf. set	Flat prior odds ratio
1	0.99	1.00	0	0.00	(A)	(A)	∞
2	0.01	1.00	0.01	1.00	(A, B)	( )	1
3	0	0.00	0.99	1.00	(B)	(B)	0

confidence intervals are redundant. It can be argued, though, that with the likelihood and a measure of overall fit both available, confidence intervals are redundant.

A. Simple Examples

Consider a simple situation where a single random variable X is observed that takes on only integer values 1, 2, or 3. It is distributed either according to model A or model B in Table I. Only $X = 2$ is compatible with both models, though it is a rare observation for either model. The implications of the model for inference are intuitively clear: If $X = 1$ is observed, model A is true, with certainty; if $X = 3$ is observed, model B is true, with certainty, and if $X = 2$ is observed, observing X has not told us anything about whether the true distribution is A or B. The odds ratios in the last column of the table capture these implications of the data precisely. The implications of observing X are well expressed by a 100% confidence interval, based on likelihood ratios, which contains A alone if $X = 1$, B alone if $X = 3$, and both A and B if $X = 2$. These are also 100% Bayesian posterior probability intervals under any prior distribution that puts nonzero weight on both models. Because the likelihood ratio (the ratio of the likelihood at a particular parameter value to the likelihood at its maximum) is always either 1 or 0 in this example, this 100% interval is the only likelihood-ratio based confidence interval available for this example.

Confidence intervals based on likelihood ratios have the advantage that they tend to have nearly the same shape as a minimum-length (or minimum-volume, in multi-parameter contexts) posterior probability interval under a flat prior. They also can be shown to have some classical optimality properties. But any collection of tests at a given exact significance level $1-\alpha$, indexed by all the points in the parameter space, implies an exact confidence region. In this example it might be appealing to test model A and model B, rejecting A if $X = 3$ and B if $X = 1$. This leads to the exact 99% confidence region listed in the seventh column of Table I. This interval agrees with the 100% interval except when $X = 2$, in which case the 99% interval is empty, instead of containing the whole parameter space.

How can we interpret a statement, having seen $X = 2$, that "with 99% confidence" the model is neither A nor B? If we seriously believe that possibly neither model is correct, this kind of statement might make sense. The confi-

TABLE II

SPURIOUSLY "INFORMATIVE" CONFIDENCE SETS

X	$P(X\|A)$	LR	$P(X\|B)$	LR	LR 99% conf. set	Flat prior odds ratio
1	0.979	1.00	0.979	1.00	(A, B)	1.00
2	0.01	0.91	0.011	1.00	(B)	0.91
3	0.011	1.00	0.01	0.91	(A)	1.10

dence interval has been constructed from likelihood ratios between each model A and B and an implicit third model—that which puts equal probability on all three values of X. This is just a way of describing what has been done here—for each parameter value (A or B) collecting the points in the sample space with lowest probability density (relative to an equal-weight probability measure over the three possible observations) to form the rejection region. The empty confidence interval when $X = 2$ is observed does suggest, if we take this implicit third model as a serious possibility, that abandoning both models A and B might be justified.

If one wants information about the possibility that no parameter value in the model or models being considered performs well, the fact that a confidence interval can provide this, and no likelihood-based interval (such as a Bayesian posterior probability interval) can do so, may be an advantage of a confidence interval. However, it is nearly universal in applied work that confidence intervals are interpreted mainly as informing us about the location of the parameter values, not about the validity of the model. An empty confidence interval is only an extreme case of a narrow confidence interval. Confidence intervals can turn out to be misleading as indicators of the precision of our knowledge about parameter location because they confound such information with information about overall model fit.

Consider the example of Table II, in which there is again a single, integer-valued observed random variable X for which there are two candidate probability models, A and B. This time, the two models are very similar, with the probabilities of each of the 3 possible values of X differing by no more than .001. A likelihood-ratio based confidence set, displayed in column 6 of the table, produces a hard-to-interpret amalgam of information about fit and information about parameter location. It is never empty, but it contains only model B when X is 2 and only model A when X is 3. A statement, on observing $X = 3$, that "with 99% confidence the model is A" makes no sense as a claim about the precision of our ability to distinguish between models A and B based on the observation. The interval is "short" because of poor overall fit of the model, not because of precision of the data's implications about choosing between A and B.

The likelihood in these simple two-model cases is information equivalent to the flat-prior odds ratio in the last column of the tables. Reporting it would correctly show for the Table II case, no matter what X were observed, that seeing X had provided little information about which model, A or B, is the truth. But if we see $X = 3$ and simply cite the odds ratio of 1.1 in favor of A, we fail to convey the fact that $X = 3$ was unlikely under either model A or B. It would be better to report the two separate p.d.f. values .011 and .01, or the odds ratio 1.1 and the marginalized likelihood (the sum of the p.d.f. over the two models, .021).

The confidence intervals in these examples give misleading information about shape of the likelihood only in low-probability samples. This reflects a general proposition, that an exact $(1-\alpha)\%$ confidence interval must have Bayesian posterior probability greater than $1 - n\alpha$ on a set of sample values with probability, under the marginal distribution over observations implied by the prior distribution and the model jointly, of at least $(n-1)/n$, for any $n > 1$. In other words, a stochastic interval with high confidence level must in most samples also have high posterior probability. The proposition is symmetric: a stochastic interval that has posterior probability $1-\alpha$ in every sample must have a coverage probability greater than or equal to $1 - n\alpha$ for a set of parameter values that has prior probability of at least $(n-1)/n$.[6]

Thus if confidence intervals or regions with confidence levels near 1 are easy to compute, they can be justified as likely to be good approximations, with high probability, to direct characterizations of likelihood shape. But where, as in the models we consider, direct characterizations of likelihood shape are easier to compute, there seems to be little reason to undertake special effort to compute confidence regions rather than direct measures of likelihood shape. Also, for characterizing likelihood shape, bands that correspond to 50% or 68% posterior probability are often more useful than 95% or 99% bands, and confidence intervals with such low coverage probabilities do not generally have posterior probabilities close to their coverage probabilities.

B. A Time Series Example

To show how these points apply in a time series model we consider the simple model

$$y(t) = \rho y(t-1) + \varepsilon(t) \qquad (t = 1, \dots, T). \tag{1}$$

We assume ε is i.i.d. $N(0,1)$, independent of y's dated earlier. Suppose a sample produces $\hat{\rho} = .95$ and $\hat{\sigma}_\rho = .046$, both likely values when $\rho = .95$. An

[6] The proof of these propositions follows from the observation that the following two things are the same: (i) the expectation over the prior distribution of the (classical) conditional coverage probabilities given parameter values, and (ii) the expectation over the marginal distribution of the data (with parameters integrated out using the prior) of the posterior probability that the parameter is in the set given the data. These are two ways to calculate the probability, under the joint probability measure on parameters and data, of the event that the parameter lies in the random set.

exact[7] finite-sample 68% confidence interval for ρ, based on the distribution of $(\rho - \hat{\rho})^2/\sigma_\rho^2$ (the log likelihood ratio), is (0.907, 0.998). A Bayesian flat-prior posterior 68% probability region is (.904, .996). Here the confidence region is not very misleading. It has flat-prior posterior probability .67 instead of .68, which is not a big error, and is shifted upward by .003 at the lower end and .002 at the upper end, which is a shift of only about 5% of its length. The likelihood levels at the upper and lower bounds of the interval differ by only about 6%.

The difference between confidence intervals and integrated-likelihood intervals becomes somewhat larger if we consider 95% intervals. For this probability level, the Bayesian interval is (0.860, 1.040), while the LR-based exact confidence interval is (0.865, 1.052). This interval shows noticeable distortion at the upper end. The likelihood is twice as high at the lower end of the interval as the upper. The interval therefore is representative of the shape of posterior beliefs only for a prior with p.d.f. increasing over the interval so that $\rho = 1.052$ is twice as likely as $\rho = .865$. The interval does have posterior probability under a flat prior of .95 (to two-figure accuracy), however.

The LR-based exact confidence interval we consider here is intuitively attractive and has some justification in classical theory, but as usual there is no unique classical confidence interval. It is commonly proposed (see Rothenberg and Stock (1997), e.g.) that intervals be based instead on the signed likelihood ratio (the signed difference in log likelihoods). Such intervals tend for this model to be more sharply asymmetric than ordinary LR intervals. In this simple case the confidence intervals based on signed LR are (0.917, 1.013) for 68% confidence and (0.869, 1.064) for 95% confidence. Though these intervals are further biased upward as likelihood descriptions than the LR-based intervals, they have flat-prior posterior possibilities of .68 and .95.

Cases of empty confidence intervals or misleadingly short confidence intervals are likely to arise when the parameter space is restricted to $\rho \in [0,1]$. If the confidence interval is based on the distribution of the signed difference between the log likelihood at ρ and the log likelihood at the unrestricted MLE $\hat{\rho}$ (which can exceed 1, of course), and is built up from equal-tailed tests, empty 68% confidence intervals have a probability of 16% and empty 95% confidence intervals have a probability of 2.5% when $\rho = 1$.[8] If instead the distribution of

[7] Since the classical small sample distribution was calculated by Monte Carlo methods on a grid with spacing .001 on the ρ-axis and with 2000 replications, "exact" should be in quotes here, but the accuracy is good enough for these examples. The method was to construct, for each ρ, 2000 artificial samples for y generated by (1) with $T = 60$, variance of ε equal to 1, and $y(0) = 0$. (The same sequence of ε's was used for all ρ's on each draw of an ε sequence.) For each ρ, an empirical distribution of the log likelihood ratio (LR) $(\hat{\rho} - \rho)^2 \cdot \Sigma y(t-1)^2$ was constructed and the 16% and 84% quantiles of the empirical distribution calculated. Then a classical 68% confidence interval can be constructed in a particular sample as the set of all ρ's for which the log LR statistic lies between the 16% and 84% quantiles appropriate for that ρ. The same method was applied in the other example calculations in this section.

[8] This is an exact result, by construction. The confidence interval lies entirely above $\rho = 1$ in samples for which $(\hat{\rho} - \rho) \cdot \sqrt{\Sigma y_{t-1}^2}$ is negative and below the lower $(\alpha/2)$% tail of its distributional conditional on $\rho = 1$. But by construction this occurs in just $(\alpha/2)$% of samples when in fact $\rho = 1$.

the ratio of the likelihood at ρ to the likelihood at its maximum on $[0,1]$ is used to generate the interval, the interval will never be empty, but it will be shorter than can be justified by the likelihood shape in these samples where the unrestricted MLE lies above 1, because of the same kind of phenomenon displayed in Table II. When the true ρ is 1 for example, an observation as high as $\hat{\rho} = 1.0387$ has about a 1% chance of occurring, and a typical associated σ_ρ is .0394. With such a sample, the exact classical 68% confidence interval based on the signed LR, calculated numerically on a grid with ρ's spaced at .001, includes only three points, defining the interval (.998, 1), and the 95% interval expands only to (.966, 1). But the likelihood has the form displayed in Figure 1, which clearly does not concentrate as sharply near 1 as these confidence intervals would suggest. Such a high $\hat{\rho}$ is unlikely for any ρ in $[0,1]$, but not as much more likely for ρ's of .97 compared to ρ's of 1 as the confidence interval would suggest. A flat-prior posterior 68% interval for ρ is (.975, 1) and the 95% probability interval is (.944, 1).

The conclusion these examples are meant to illustrate is that flat-prior Bayesian $1-\alpha$ probability intervals, based purely on the likelihood, are often very close to exact $(1-\alpha)$% confidence intervals, and that in those samples where they differ, confidence intervals can be quite misleading as characterizations of information in the sample about parameter location.

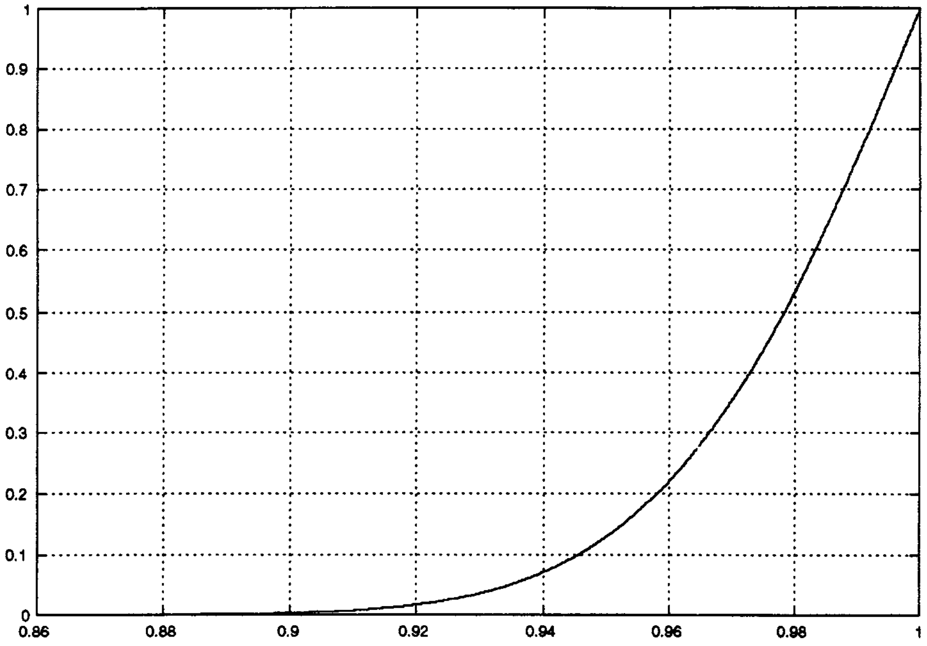

FIGURE 1.—Likelihood over $(0,1)$ with $\hat{\rho} = 1.0387$, $\sigma_\rho = .0394$.

3. MULTIVARIATE DYNAMIC MODELS

Consider a model of the form

$$g(y(t), y(t-1), \dots, y(t-k) | \beta) = \varepsilon(t) \tag{2}$$

where

$$\left| \frac{\partial g(y(t), y(t-1), \dots, y(t-k) | \beta)}{\partial y(t)} \right| \neq 0, \tag{3}$$

so that the equation defines $y(t)$ as a function of $\varepsilon(t), y(t-1), \dots, y(t-k), \beta$. We assume that $\varepsilon(t)$ is independent of $y(t-s)$, all $s > 0$, and has p.d.f. $h(\cdot | \Omega)$. Under these conditions (3) allows us to write a p.d.f. for $y(1), \dots, y(T)$ conditional on $y(0), \dots, y(-k+1), \beta, \Omega$, which we will label

$$p(y(1), \dots, y(T) | y(0), \dots, y(-k+1), \beta, \Omega). \tag{4}$$

It is common but not universal practice to treat (4) as the likelihood function. It is actually the likelihood function only if the distribution of $y(0), \dots, y(-k+1)$ does not depend on unknown parameters, or depends only on unknown parameters unrelated to β and Ω. If (2) is consistent with y being ergodic, then it is natural to take the ergodic marginal distribution of $y(t), \dots, y(t-k)$, which of course in general depends on β and Ω, as the distribution for $y(0), \dots, y(-k+1)$. This is then combined with (4) to produce a p.d.f. for the full vector of observations $y(-k+1), y(-k+2), \dots, y(T)$. There are three reasons this is not often done: it often makes computations much more difficult; it requires that we rule out, or treat as a separate special case, nonstationary (and thus nonergodic) versions of the model; and it may not be plausible that the dynamic mechanism described in (2) has been in operation long enough, in unchanged form, to give $y(0), \dots, y(-k+1)$ the ergodic distribution. The last two points are related. A nonstationary model has no ergodic distribution. A near-nonstationary model may have an ergodic distribution, yet it may imply that the time required to arrive at the ergodic distribution from arbitrary initial conditions is so long that imposing the ergodic distribution on $y(0), \dots, y(-k+1)$ may be unreasonable.

Bayesian inference would ideally use the ergodic distribution for the initial conditions at parameter values well within the stationary region of the parameter space, then shift smoothly over to a distribution less connected to β and Ω as the nonstationary region is approached. Such a model is likely to be application-dependent, however, and in the remainder of this paper we treat (4) as the likelihood. We also hold initial conditions fixed in generating Monte Carlo samples of $y(1), \dots, y(T)$ when we evaluate classical convergence probabilities.[9]

[9] On this latter point we differ from Kilian (1998a). Conditioning on initial data values in forming the likelihood or in constructing estimators (on which point Kilian's practice matches ours) amounts to ignoring potential information on the initial observations. Conditioning them in doing Monte Carlo simulations of the data-generation process amounts to recognizing the distinction between initial conditions that generate more and less informative samples. If the initial y's show unusually large deviations from their steady-state values, the sample is likely to generate unusually sharp information about the parameters. It does not make sense to calculate coverage probabilities that average across informative and uninformative samples when it is easy to take account of the fact that we have been lucky (or unlucky) in the initial conditions of the particular sample at hand.

Much of our analysis will focus on the case where g is linear in its first y argument with an identity coefficient matrix and also linear in β, and h is Gaussian. Under these conditions the log likelihood is quadratic in β, so the likelihood itself is Gaussian in shape, in small and large samples, for stationary and nonstationary models.

For a general model of the form (2), there is ambiguity about how "impulse responses" ought to be defined. Here, though, we consider only models such that g is linear in its y arguments, so that the response $c_{ij}(s)$ of $y_i(t+s)$ to $\varepsilon_j(t)$ is easily and unambiguously defined. By solving (2) recursively, we can solve for $y(t+s)$ as a function of $\varepsilon(t+s), \varepsilon(t+s-1), \ldots, \varepsilon(t)$ and $y(t-1), \ldots, y(t-k)$. Then

$$c_{ij}(s) = \frac{\partial y_i(t+s)}{\partial \varepsilon_j(t)} \tag{5}$$

depends only on β, not y or ε, and can be calculated by elementary matrix operations. Note that, though the impulse responses as defined here coincide with the coefficients of the moving average representation (MAR) for stationary, linearly regular models, they are also defined for nonstationary models where no MAR exists.

If we write the linear model in terms of the lag operator, as

$$A(L)y(t) = \varepsilon(t), \tag{6}$$

then the impulse responses are the coefficients in the infinite-order polynomial in the lag operator $C(L) = A^{-1}(L)$, where the inverse is restricted to involve only positive powers of L and is interpreted as an operator on the field of finite-order polynomials in the lag operator. Such an inverse exists for any finite-order A with A_0 full rank, regardless of the characteristic roots of A, though of course it may imply coefficient matrices C_s that fail to converge to zero as $s \to \infty$. The mapping from the coefficients in A, the autoregressive form of the model, to those in C, is one-one. Because A is finite-order (in the models we consider) and C is not, there is a sense in which A is a more economical characterization of the model than is C. On the other hand, because the elements $c_{ij}(t)$ of the C_t matrices tend to behave, as function of t, like the data that is being modeled, the properties of the model are often more easily grasped intuitively by viewing plots of $c_{ij}(t)$'s than by viewing tables of $a_{ij}(t)$'s.

Though the mapping from the full A operator to the full C operator is one-one, the mapping from the a_{ij}'s to a particular $c_{ij}(t)$ is complicated. Generally the mapping is not one-one, even if we consider a collection $c_{ij}(t)$, $t = 1, \ldots, H$, in which H is the same as the dimension of the vector of all the free parameters in A.[10] This engenders some complications in defining error bands, and (especially) in defining confidence intervals, for $c_{ij}(t)$.

[10] To see this point, consider a first-order model with $A = I - A_1 L$ and A_1 $m \times m$. If A_1 has Jordan decomposition $A_1 = P\Lambda P^{-1}$, with Λ diagonal, then $c_{ij}(\cdot)$ depends on the ith row of P, the jth column of P^{-1}, and all the elements of the diagonal of Λ, but on no other aspects of A_1. Clearly this leaves room for many A_1's to deliver exactly the same $c_{ij}(\cdot)$.

4. DIMENSIONALITY

Because impulse responses are high dimensional objects, describing error bands for them is challenging. The best approaches to describing them will differ to some extent across applications, but we describe some widely applicable approaches in Section 6 below. In this section we discuss the difficulties created by the fact that multivariate time series models have high-dimensional parameter spaces.

A complete characterization of a subset of a parameter space in a 150-dimensional parameter space (which would not be unusually large for a multivariate time series model) would be too complicated to be useful. Instead we try to capture substantively important characteristics of error band sets with a few lower-dimensional statistics and plots. For example, we may focus attention on a few impulse responses that are important for characterizing the effects of a policy action on the economy. With a given Bayesian prior distribution, characterizing the probability, given the data, of a set of possible values for a given impulse response—e.g. the probability that $c_{ij}(t) \in [a, b]$, $t = 1, \ldots, 4$—is a straightforward exercise: One integrates the posterior p.d.f. with respect to all the parameters of the model over the set of parameter values that deliver c_{ij}'s satisfying the restrictions.

For confidence intervals, the high dimensional parameter space raises a different kind of problem. The same impulse response function c_{ij} can arise from two different $A(L)$ operators. The coverage probability of any stochastic set in the space of c_{ij}'s, no matter how generated, will depend not just on c_{ij} itself, and will therefore be different for different $A(L)$'s corresponding to the same c_{ij}. Therefore no exact confidence set for c_{ij} is possible.

This point, that coverage probabilities often depend on nuisance parameters and that this causes problems, is an old one. Some writers define the confidence level of a random interval as the minimum over the parameter space of the coverage probability.[11] This corresponds (because of the duality between confidence regions and hypothesis tests) to the standard definition of the size of a statistical test as the maximum over the parameter space of the probability of rejection under the null. However, as has recently been emphasized in work by Dufour (1997), Faust (1996), and Horowitz and Savin (1998), this approach is unsatisfactory, as it leads often to a situation where there are no nontrivial confidence intervals, even asymptotically. In practice, researchers use confidence intervals or regions justified by asymptotic theory, in the sense that the coverage probability converges to its nominal level as sample size increases, pointwise at each point in the sample space. Often this can be accomplished by simply computing coverage probabilities at estimated values of nuisance parameters.

In small samples, though, this approach can lead to serious distortion of the shape of the likelihood. For example, there can be a wide range of values of a

[11] Zacks (1971), e.g. defines confidence level this way, while Wilks (1962) seems to consider in small samples only cases of coverage probabilities uniform over the whole parameter space.

nuisance parameter υ, all with high likelihood, with widely different implications for the coverage probability of a confidence set. If we replace υ by a single estimated value, we may obtain an unrepresentative value for the coverage probability. In the same situation, Bayesian posterior probability intervals would involve integration over υ, rather than choice of a single υ, which better represents the actual implications of the data. With the likelihood spread evenly over υ's with different implications, Bayesian results could be sensitive to the prior. But a thorough description of the likelihood, including if necessary consideration of more than one reference prior, can make this situation clear. Furthermore, often the likelihood dominates the prior, so that over the range of υ's of interest most readers would agree that the prior is nearly flat.

5. ASYMPTOTICS APPROXIMATION, THE BOOTSTRAP

We have already cited the fact that Bayesian posterior probability intervals have asymptotic justification as confidence intervals in time series models of the type we consider. This result, like many on which confidence intervals for these models have been based, actually applies only to situations where the model is known to be stationary. The result is easiest to see in the special case of an unrestricted reduced form VAR, i.e. a model with $A_0 = I$. In that case the likelihood conditioned on initial observations is Normal-Inverse-Gamma in shape, centered at the OLS estimate, with a covariance matrix that even in finite samples is exactly the covariance matrix of the asymptotic distribution of the estimates. The marginal distribution for the A_j's is then easily seen to converge to the shape of the same limiting normal distribution to which the sampling distribution of the OLS estimates converges. Of course this then implies that all differentiable functions of the coefficients of A, including the C_j's, have posterior distributions that are asymptotically normal, and match the sampling distributions of the same functions of the OLS estimate of A.

For more general stationary linear models a maximum likelihood estimator of the coefficients in A, or a pseudo-maximum likelihood estimator based on a Gaussian likelihood when the disturbances are not Gaussian, or another GMM estimator, will each converge, under mild regularity conditions, to a joint normal distribution. The posterior distribution of the parameters, conditional on the estimators themselves and their estimated asymptotic covariance matrix (not conditional on the full sample, outside the Gaussian case), should in large samples be well approximated by taking the product of the prior p.d.f. with the asymptotic normal p.d.f. as an approximate joint p.d.f. from which to construct the posterior. But since in large samples the likelihood dominates the prior, we can treat the prior as constant in this approximation. The computation of the posterior amounts to fixing the estimator and its estimated covariance matrix, then treating the asymptotic p.d.f. of the sample as a function of the coefficients in A in constructing the posterior p.d.f. for A. This will make the posterior distribution have the same shape as the asymptotic sampling distribution of the estimator. Because the asymptotic distribution is symmetric and affected by the parameters only via a pure location shift, confidence regions then coincide with posterior probability intervals. Proofs of these results, with careful specification of regularity conditions, are in Kwan (1998).

The mappings from A_j coefficients to C_j coefficients can be analytically differentiated, so that normal asymptotic distributions for the A coefficients can be translated into normal asymptotic distributions for the C coefficients. Details have been provided, e.g., in Lutkepohl (1990) and Mittnik and Zadrozny (1993).[12] However the mapping from A to C_j is increasingly nonlinear as j increases, so that the quality of the approximation provided by this type of asymptotic theory deteriorates steadily with increasing j in any given sample. Eventually as j increases (and this occurs for fairly low j in practice) the norm of the variance of C_j begins to behave as λ^{-j}, where λ is the root of $|A(L)|=0$ that is smallest in absolute value. This means that when the point estimate of λ exceeds one, error bands shrink at the rate λ^{-j} for large j, even when there is in fact substantial uncertainty about the magnitude of λ. This is an important source of inaccuracy, and results in poor behavior in Monte Carlo studies of error bands generated this way. (See Kilian (1998b).)

Another way to generate confidence intervals with the same first-order asymptotic justification as use of Bayesian intervals as confidence intervals is to use simple bootstrap procedures. Perhaps because Bose (1988) has shown that bootstrap calculations provide a higher order of asymptotic approximation in the distribution of the estimator in a stationary finite-order autoregressive model, some researchers have the impression that simple methods for translating an estimate of the distribution of the estimator at the true parameter value into a confidence interval must improve the accuracy of confidence intervals, but this is not the case. Some applied studies[13] have used what might be called the ''naive bootstrap,'' but which Hall (1992) and we in the rest of this paper will call the ''other-percentile'' bootstrap interval. An initial consistent estimate $\hat{A}$ is generated, and Monte Carlo draws are made from the model with $A=\hat{A}$, to generate a distribution of estimates $\hat{\hat{A}}$ about $\hat{A}$. To generate a confidence band for $c_{ij}(t)$, the $\hat{\hat{A}}$'s are mapped into a distribution of corresponding $\hat{c}_{ij}(t)$'s, and the $(1-\alpha)\%$ ''confidence interval'' is formed by finding the upper and lower $(\alpha/2)\%$ tails of the distribution of the $\hat{c}_{ij}(t)$'s. This procedure clearly amplifies any bias present in the estimation procedure. In our simple time series example of Section 2B, the procedure would produce intervals shifted *downward* relative to the Bayesian intervals, instead of the slight upward shift produced by exact classical intervals.

In a pure location problem, where we observe $X \sim f(X-\mu)$ and the p.d.f. f is asymmetric, the standard method of generating a confidence interval is to find the upper and lower $(\alpha/2)\%$ tails of the p.d.f. f, say a and b, and then use the

[12] This approach has been used by, e.g., Poterba, Rotemberg, and Summers (1986).

[13] This approach has been used by Runkle (1987). Blanchard and Quah (1989) and Lastrapes and Selgin (1994) use a modification of it that makes ad hoc adjustments to prevent the computed bands from failing to include the point estimates.

fact that, with μ fixed and X varying randomly,

$$P[a < X - \mu < b] = P[X - b < \mu < X - a] = 1 - \alpha. \tag{7}$$

If we used the other-percentile bootstrap and made enough Monte Carlo draws to determine the shape of f exactly, we would use not the $(X - b, X - a)$ interval implied by (7), but instead $(X + a, X + b)$—that is, the interval in (7) "flipped" around the observed point X. It may seem obvious that, having made our bootstrap Monte Carlo draws, we should use the interval in (7), not the other-percentile bootstrap interval. Indeed, the interval $(X - b, X - a)$ implied by (7) is called the "percentile interval" by Hall (1992) and treated by him as the most natural bootstrap interval. In the pure location-shift context, it produces an interval with flat-prior Bayesian posterior probability equal to its nominal coverage probability, and is therefore a useful descriptor of the likelihood function.

But we can also imagine a setting in which we do not have a location shift problem with non-Gaussian distribution, in which instead there is an underlying Gaussian (or otherwise symmetrically distributed) location-shift problem, that has become non-Gaussian through an unknown nonlinear transformation. That is, we observe X and wish to form a confidence interval for μ, as before, but now $X = g(Z)$ and $\mu = g(\theta)$, where $Z \sim N(\theta, \sigma^2)$, g is monotone, and we do not necessarily know g. It can be shown that in this situation the other-percentile bootstrap interval would give exactly the right answer, in the sense that it would produce an interval with flat-prior Bayesian posterior probability matching its nominal coverage probability.

In time series problems, for parameters that imply stationary data, estimates of the coefficients in $A(L)$ are asymptotically normal, but biased in small samples. The coefficients in C_j are functions of the coefficients in $A(L)$, with the degree of nonlinearity in the mapping increasing with j. One might expect then that bias considerations could dominate for j near zero, making Hall's percentile intervals work best, while the effects of nonlinear transformation would dominate for large j, making "other-percentile" intervals work best. But there is no obvious, general method of using bootstrap simulations to produce confidence intervals whose length and skewness accurately reflect asymmetry in sample information about the C_j's.

Kilian in several papers (Kilian (1998a, 1998b), e.g.) has argued for use of other-percentile bootstrap intervals with bias correction.[14] While there is no argument available that such intervals are more accurate in coverage probability asymptotically than other bootstrap approaches or than Bayesian probability intervals, the bias-correction does tend to remove the most important source of bad behavior of other-percentile intervals in time series models.

[14] But note that Kilian is primarily just correcting the initial estimator for bias. See Hall (1992, p. 129) for some elaboration of the point that what is called "bias correction" in the literature of bootstrapped confidence intervals is not correction of the bootstrapped estimator for bias.

There are ways to construct bootstrap confidence intervals that make them asymptotically accurate to second order, instead of only to first order. One such method is that presented in Hall (1992) as ''symmetric percentile-t.'' Another is that presented in Efrom and Tibshirani (1993) as ''BC_{α}.''[15] The Hall intervals are symmetric about point estimates, so they cannot provide information about asymmetry in the likelihood. The Efron and Tsibshirani estimates are burdensome to compute. Both have behaved badly in Monte Carlo studies of multivariate time series models (Kilian (1998b)).

As should already be clear, we believe error bands that are descriptive of the likelihood are to be preferred, in scientific reporting, to classical confidence regions with known coverage probabilities, even when the latter can be constructed. Nonetheless we report in our examples below some measures of the behavior of bootstrap-based approximate confidence intervals. We use Kilian's bias-adjusted bootstrap method (Kilian (1998a)) and the commonly applied other-percentile method. As we have noted, both can be justified as good approximations to Bayesian intervals under particular assumptions on the way the distribution of the observed statistics depends on the parameter. Both are relatively straightforward to compute.

6. BETTER MEASURES OF UNCERTAINTY ABOUT SHAPE OF RESPONSES

Common practice is to compute a one-dimensional error band $\hat{c}_{ij}(t) \pm \delta_{ij}(t)$ for each $t = 0, \ldots, H$, then plot on a single set of axes the three functions of time $\hat{c}_{ij}(t) - \delta_{ij}(t)$, $\hat{c}_{ij}(t)$, and $\hat{c}_{ij}(t) + \delta_{ij}(t)$, attempting to show the reader both the point estimate of the response and an indication of the range of uncertainty about the form of the response. It is extremely unlikely in economic applications that the uncertainty about $c_{ij}(t)$ is independent across t, and indeed this is probably a good thing. Readers probably tend to think of the plotted upper and lower confidence band points as representing $c_{ij}(\cdot)$ functions at the boundaries of likely variation in c_{ij}. If the uncertainty about c were in fact serially independent across t, the smoothly connected upper and lower confidence bands would represent extremely unlikely patterns of deviation of c_{ij} from $\hat{c}_{ij}$. But in models fit to levels of typically serially correlated economic time series, uncertainty about c is usually itself positively serially correlated, so thinking of the $\hat{c}_{ij}(t) \pm \delta_{ij}(t)$ plots as possible draws from the distribution of c is not as unreasonable as it might seem.

Nonetheless, models are sometimes fit to data that is not very smooth, or to differenced data, and even for data in levels the ''connect the dots'' bands can turn out to be misleading. To better characterize the uncertainty, one can turn to representing the $c_{ij}(\cdot)$ functions in a better coordinate system than the vector of their values at $t = 0, \ldots, H$. If the $\{c_{ij}(t)\}_{t=0}^{H}$ vector were jointly normal with covariance matrix Ω, the obvious coordinate system would be formed by

[15] This interval is an other-percentile interval adjusted for ''bias'' and ''acceleration.''

projections on the principal components of Ω, and one can use this coordinate system even if the c's are not jointly normal.

To implement this idea, one computes a pseudo-random sample from the distribution of c_{ij}, using it to accumulate a first and second moment matrix. From these, one computes the estimated covariance matrix Ω of the H-dimensional c_{ij} matrix and calculates its eigenvector decomposition

$$W\Lambda W' = \Omega, \tag{8}$$

where Λ is diagonal and $W'W = I$. Any c_{ij} can now be represented as

$$c_{ij} = \hat{c}_{ij} + \sum_{k=1}^{H} \gamma_k W_{\cdot k}, \tag{9}$$

where we are treating c_{ij} as an H-dimensional column vector, $\hat{c}_{ij}$ is the estimated mean of c_{ij}, and $W_{\cdot k}$ is the kth column (i.e. kth eigenvector) of W. Random variation in c_{ij}, which represents our uncertainty about c_{ij} given observation of the data, is generated by randomness in the coefficients $\{\gamma_k\}$. By construction, the variance of γ_k is the kth diagonal element of matrix Λ in (8), i.e. the kth eigenvalue of Ω.

In models of economic data, it is likely that the largest few eigenvalues of Ω account for most of the uncertainty in c_{ij}. We can, in such a case, give a much more complete description of uncertainty about c_{ij} than is available in a "connect-the-dots" error band by presenting just a few plots, displaying, say, $\hat{c}_{ij}(t) \pm W_{\cdot k}(t) \cdot \sqrt{\lambda_k}$ to represent an approximate 68% probability band for the kth variance component, or $\hat{c}_{ij}(t) \pm W_{\cdot k}(t) \cdot 1.96\sqrt{\lambda_k}$ for approximate 95% intervals. Unlike the functions of t plotted in connect-the-dots error bands, these plots show c_{ij} functions that lie in the boundary of the Gaussian confidence ellipsoid, when the distribution of c_{ij} is Gaussian.

This method, though possibly useful in practice as a time-saver, is still unsatisfactory, as we know that the c_{ij}'s are often not well approximated as jointly normal, particularly for their values at large t. The linear eigenvector bands we have described will always be symmetric about $\hat{c}_{ij}$, for example, while the posterior distribution usually is not. It is often important to know when the posterior is strongly asymmetric.

So we can improve on the above method by conducting another round of Monte Carlo simulation, or, if the first round, used to form Ω, has been saved, by making another pass through the saved draws. This time, we tabulate the 16%, 84%, 2.5% and 97.5% (say) quantiles of the γ_k corresponding to the largest eigenvalues of Ω. The γ_k for a particular draw of c_{ij} is easily computed as $W_{k\cdot}c_{ij}$, where $W_{k\cdot}$ is the kth row of W. The two functions of t, $\hat{c}_{ij} + \gamma_{k,.16}$ and $\hat{c}_{ij} + \gamma_{k,.84}$, if λ_k is one of the largest eigenvalues, will each show a likely direction of variation in c_{ij}, and their different magnitudes of deviation from $\hat{c}_{ij}$ will give an indication of asymmetry in the likelihood or posterior p.d.f.

One further step up in sophistication for these measures is possible. Often it is important to the use or the interpretation of a result how several c_{ij}'s, for different i and j, are related to each other. For example, it is often thought that plausible patterns of response of the economy to a disturbance in monetary policy should have interest rates rising, money stock falling, output falling, and prices falling. When we look at the error band for the response of output to a monetary shock and see that, say, it looks plausible that the output decline could be much greater than the point estimate, we really need to know whether the stronger output decline implies any violation of the sign pattern that implies the shock is plausibly treated as a policy shock.

This kind of pattern can be answered by stacking up several c_{ij}'s into a single vector, then applying eigenvalue decomposition methods suggested above. It may not be as widely useful to apply these methods to c_{ij}'s jointly as it is to apply them to single c_{ij}'s, because there is not as strong a presumption of high correlations of $c_{ij}(t)$ values across i and j as there is across t. However, where it turns out that a few eigenvalues do dominate the covariance matrix of a substantively interrelated set of c_{ij}'s, this sort of exercise can be very helpful.

Before proceeding to examples, we should note that the ideas in this section are not limited in application to impulse responses. They apply wherever there is a need to characterize uncertainty about estimated functions of an index t and there is strong dependence across t in the uncertainty about the function values. This certainly is the case when we need to characterize uncertainty about the future time path of an economic variable in a forecasting application, for example. It also will apply to some cases in which nonparametric kernel estimates of regression functions, p.d.f.'s, or the like are being presented.

7. APPLICATION TO SOME BIVARIATE MODELS

We consider two examples of bivariate models based on economic time series. Though they are unrealistically simple, they make it feasible to do a wide range of checks on the methods, in order to give a clearer impression of the feasibility of the methods we suggest, their value in transmitting information, and the degree to which they give results resembling or differing from other possible methods for generating error bands.

First we consider a triangularly orthogonalized reduced-form VAR with four lags fitted to quarterly data on real GNP and $M1$ over 1948:1–1989:3. Figure 2 shows pointwise flat-prior posterior probability intervals for this model's impulse responses.[16] For both the response of Y to its own shocks, in the upper left, and the response of $M1$ to its own shocks, in the lower right, there is notable skewness in the bands, with the 95% bands especially extending farther up than down. Figure 3 shows 68% other-percentile intervals, without bias-correction. Note that for the responses of GNP and $M1$ to their own innovations, these

[16] For both these models we use 1000 Monte Carlo draws in generating Bayesian or bootstrap intervals for a given data set, and 600 draws of data sets in constructing coverage probabilities.

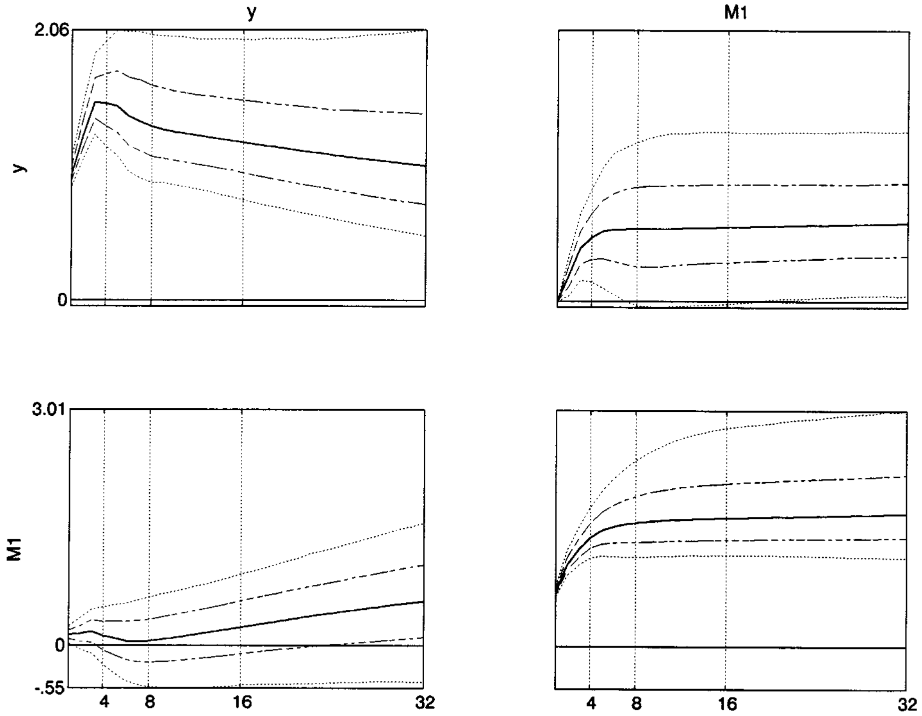

FIGURE 2.—Pointwise .68 and .95 posterior probability bands, Y-M Model.

other-percentile bands are quite asymmetric, shifted down toward 0. This reflects bias in the estimator, and it does not make sense to treat these intervals, which accurately reflect the shape of the distribution of the estimates about the truth, as characterizing reasonable beliefs about parameter location. With bias-adjustment, as displayed in Figure 4, other-percentile bootstrap intervals show the same qualitative pattern as the Bayesian intervals, but are more skewed, somewhat wider, and with a tendency to splay out at the longer time horizons.[17]

Table III, computed for the posterior implied by the actual data, shows that the bias-corrected other-percentile bootstrap intervals have posterior probability that substantially exceeds .68. This results from an even greater excess length for these intervals, as indicated by Table IV. Treating the point estimates of this model as true parameter values, we can examine coverage probabilities. As can be seen in Table V, Bayesian .68 intervals have classical coverage probability

[17] Note that our implementation of bias-adjusted other-percentile intervals does not follow Kilian in his ad hoc adjustments of sample draws to eliminate nonstationary roots. Our view is that mildly nonstationary estimates may be realistic, so that it is usually inappropriate to truncate the parameter space at or just inside the boundary of the stationary region. We also condition on initial observations in generating bootstrap draws, which Kilian does not.

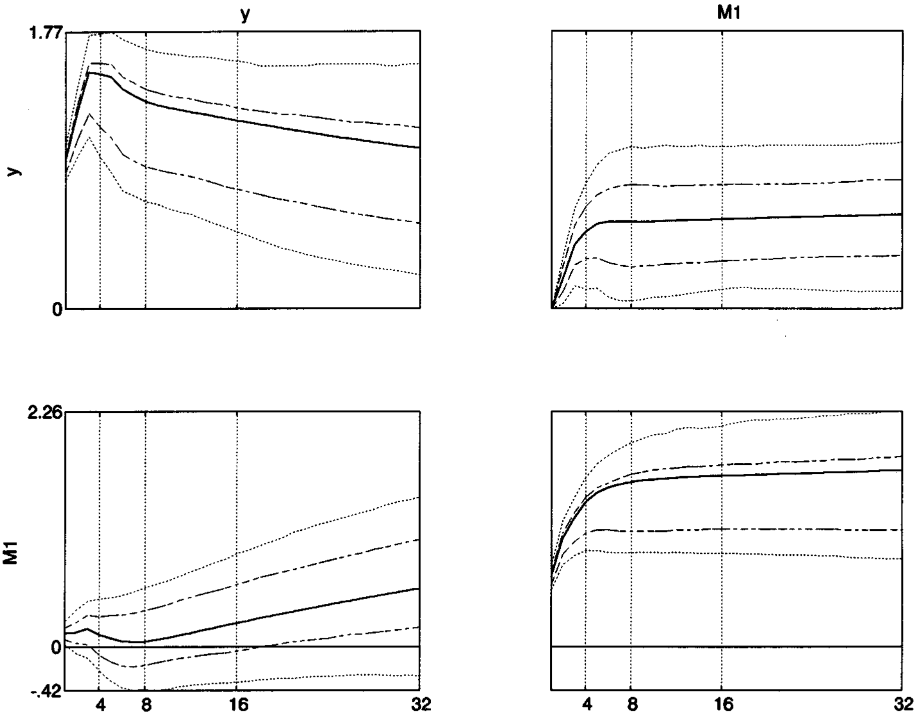

FIGURE 3.—68% and 98% other-percentile bootstrap bands, Y-M Model.

TABLE III

BOOTSTRAP INTERVALS AS LIKELIHOOD DESCRIPTORS, *GNP-M*1 MODEL

	Posterior Probability of Bootstrap Interval			
t	*Y* to *Y*	*M* to *Y*	*Y* to *M*	*M* to *M*
1	.613	.685	.000	.809
2	.723	.722	.680	.799
3	.716	.697	.683	.780
4	.742	.713	.678	.760
6	.729	.751	.732	.725
8	.726	.769	.765	.702
12	.778	.790	.818	.671
16	.788	.822	.841	.675
24	.801	.831	.868	.716
32	.809	.832	.883	.776

Note: Monte Carlo standard error of a .68 frequency with 1000 draws is .015. Bootstrap intervals have nominal coverage probability of .68. The samples are random draws with initial *y*'s both 1.

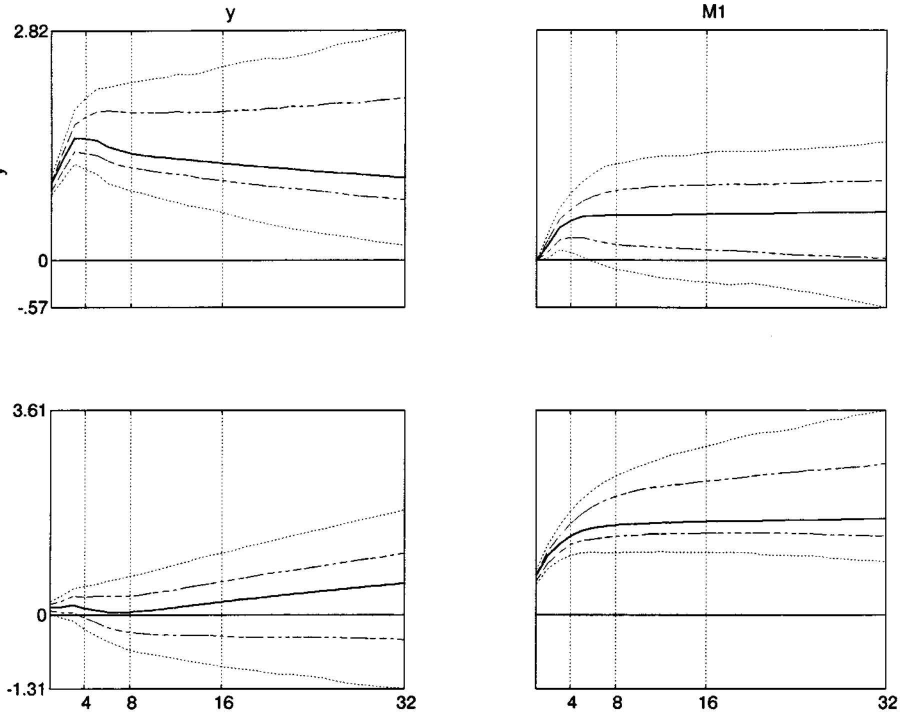

FIGURE 4.—68% and 95% bias-corrected bootstrap bands, Y-M Model.

TABLE IV

COMPARISON OF INTERVAL LENGTHS, *GNP-M*1 MODEL

	Ratios of Mean Lengths: Bayesian/Bootstrap			
t	*Y* to *Y*	*M* to *Y*	*Y* to *M*	*M* to *M*
1	1.022	1.016	1.000	1.020
2	.995	.988	1.015	.983
3	.967	.961	.982	.949
4	.938	.929	.945	.918
6	.895	.880	.892	.872
8	.841	.829	.836	.828
12	.759	.762	.750	.756
16	.716	.730	.708	.721
24	.652	.691	.644	.677
32	.603	.661	.599	.647

TABLE V

BAYESIAN AND BOOTSTRAP INTERVALS AS CONFIDENCE REGIONS, *GNP-M*1 MODEL

Bootstrap Interval Coverage Probabilities					Bayesian Interval Coverage Probabilities			
Y to *Y*	*M* to *Y*	*Y* to *M*	*M* to *M*	*t*	*Y* to *Y*	*M* to *Y*	*Y* to *M*	*M* to *M*
.432	.687	.000	.360	1	.595	.668	.000	.595
.540	.678	.717	.502	2	.618	.672	.685	.543
.603	.692	.685	.592	3	.577	.687	.643	.557
.625	.670	.688	.622	4	.593	.690	.613	.518
.657	.702	.687	.653	6	.585	.677	.618	.523
.650	.713	.685	.655	8	.560	.670	.610	.510
.682	.717	.697	.675	12	.555	.642	.627	.518
.703	.732	.707	.687	16	.557	.623	.608	.513
.712	.730	.742	.712	24	.570	.642	.617	.517
.725	.742	.748	.750	32	.602	.632	.610	.520

Note: Monte Carlo standard error of a .68 frequency with 600 draws is .019. Bootstrap intervals have nominal coverage probability of .68. Bayesian intervals are flat-prior, equal-tail, .68 posterior probability intervals.

systematically lower than .68, though never lower than .50. The bootstrap intervals have coverage probabilities less than .68 for short horizons, above .68 for long horizons, but the discrepancies are smaller than for the Bayesian intervals at most horizons.

For this model, the pointwise bands carry nearly all shape information, because the first component of the covariance matrix of each impulse response is strongly dominant. We can see this in Figure 5, which displays error bands for the first component of variation in each impulse response. The bands in this figure are very similar to the pointwise bands shown in Figure 2. The main difference is a tendency for the component bands to be tighter for t close to 0. For each of the four response functions in these graphs, the largest eigenvalue accounts for over 90% of the sum of the eigenvalues. The second eigenvalue accounts for between 2.3% and 32%, and the third for no more than 1.1%. Figure 6 shows the second component, which in this case seems to correspond to uncertainty about the degree to which the response is “hump-shaped.” We do not display the third component because it is quite small. In this model, use of pointwise bands, together with what we guess is the usual intuition that uncertainty about the level, not the shape, of the response dominates, would lead to correct conclusions.

We turn now to the bivariate VAR with 8 lags fitted by Blanchard and Quah (1989) to quarterly data on real GNP growth and the unemployment rate for males 20 and over, for 1948:1–1987:4. We construct error bands for the structural impulse responses under the Blanchard-Quah identification, which relies on long-run restrictions. The original article presented what were meant to be other-percentile intervals, with an ad hoc correction to prevent their lying

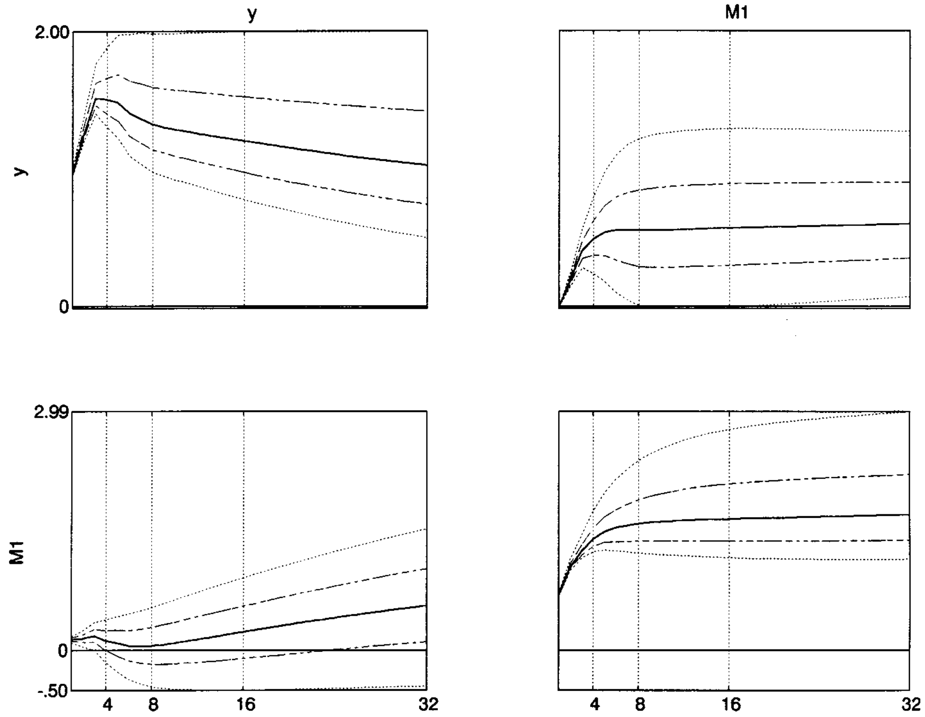

FIGURE 5.—First component: .68 and .95 probability bands, Y-M Model.

entirely to one side of the estimated responses.[18] We were able to duplicate those results, but realized in the process that there were some errors in the paper's implementation of the bootstrap. The very strong asymmetry in the intervals displayed in the article resulted mostly from the errors.[19] Figure 7 shows Bayesian flat-prior intervals. They are only modestly asymmetric, but they are rather wide for the responses to a supply shock. The uncorrected other-percentile intervals in Figure 8 are similar to the Bayesian intervals, except for strong skewness for low *t* in the responses to demand shocks. The first few periods show 68% bands lying entirely below the point estimates, and we know from Figure 7 that this does not reflect the shape of the likelihood. Bias correction has slight effects on these bands, as can be seen from Figure 9; in particular the strong bias toward zero in the bands for responses to demand at low *t* is still present after bias-correction.

[18] This same general procedure was followed also by Lastrapes and Selgin (1994). For the Blanchard-Quah model, Koop (1992) computed Bayesian error bands and noted how different they were from the original (numerically mistaken) published bands.

[19] See the more detailed discussion in an earlier draft of this paper, Cowles Foundation Discussion Paper Number 1085.

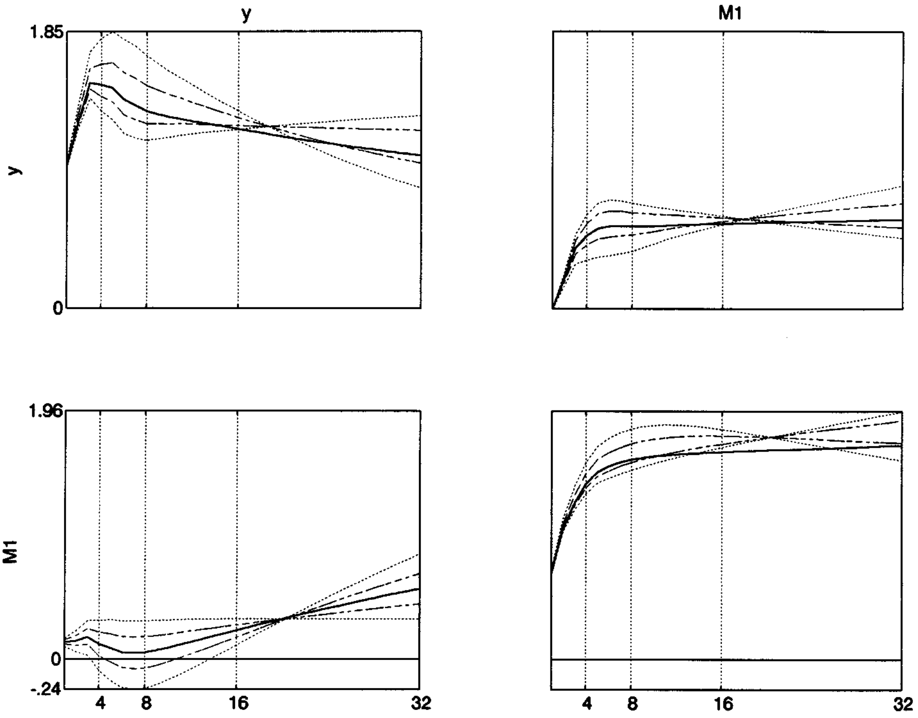

FIGURE 6.—Second component: .68 and .95 probability bands, Y-M Model.

In contrast to the implications of Table III for the *Y-M* model, Table VI shows that the bias-corrected bootstrap intervals are for the B-Q model not bad as indicators of posterior probability, except for a tendency to hold too little probability at low *t* in the responses to demand shocks. This reflects the spurious skewness already noted for those responses and *t* values. In this model, in contrast to the *Y-M*1 model, the Bayesian intervals and bias-corrected bootstrap intervals have almost the same average lengths, as can be seen from Table VII. The lower posterior probability for the bootstrap intervals at low *t* for responses to demand comes from the intervals being mislocated, not from their being too short.

If we take the estimated coefficients of the Blanchard-Quah model as the truth, we can check coverage probabilities for bootstrap and Bayesian intervals. Table VIII shows similar behavior for Bayesian and bootstrap intervals by classical criteria. Both intervals tend to undercover for the *GNP*-to-demand response at short horizons, though the undercoverage is considerably worse for the bootstrap interval. Both tend to overcover at long time horizons, with the tendency slightly worse for the Bayesian intervals.

For this model, the first component of the covariance matrix accounts for much less of the overall variation, especially for the responses to demand, as is

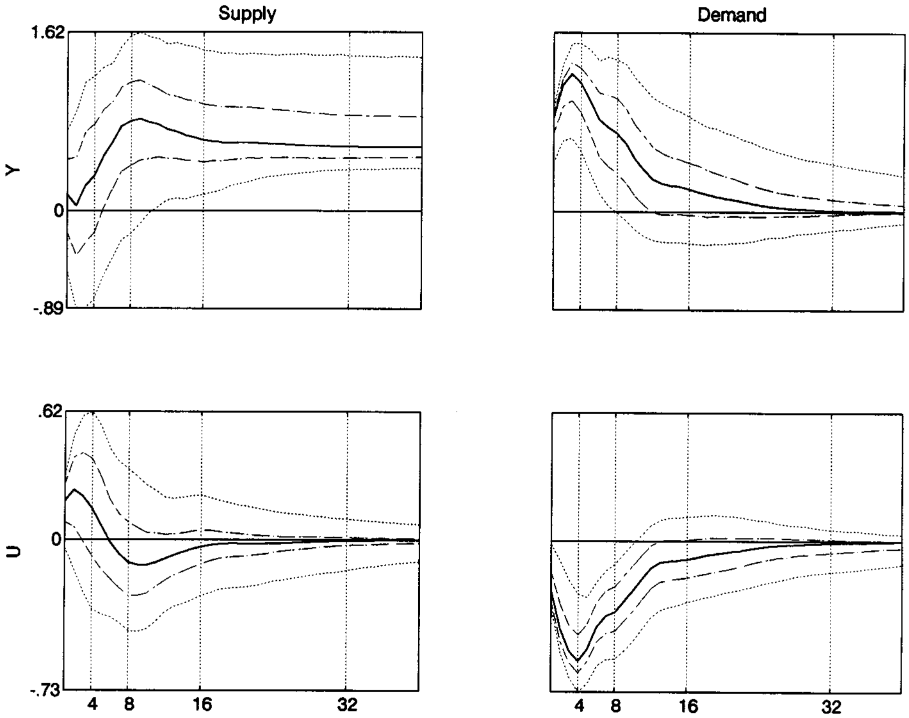

FIGURE 7.—Pointwise: .68 and .95 posterior probability bands, B-Q Model.

TABLE VI

BOOTSTRAP INTERVALS AS LIKELIHOOD DESCRIPTORS, BLANCHARD-QUAH MODEL

	Posterior Probabilities of Bootstrap Intervals			
t	*Y* to *S*	*U* to *S*	*Y* to *D*	*U* to *D*
1	.710	.696	.452	.644
2	.706	.715	.583	.616
3	.701	.717	.655	.611
4	.688	.712	.700	.661
6	.693	.696	.712	.673
8	.703	.698	.705	.731
12	.687	.718	.700	.692
16	.693	.713	.689	.684
24	.668	.714	.670	.676
32	.655	.721	.707	.718

Note: Monte Carlo standard error of a .68 frequency with 1000 draws is .015. Bootstrap intervals have nominal coverage probability of .68. The samples are random draws with initial *y*'s both 1.

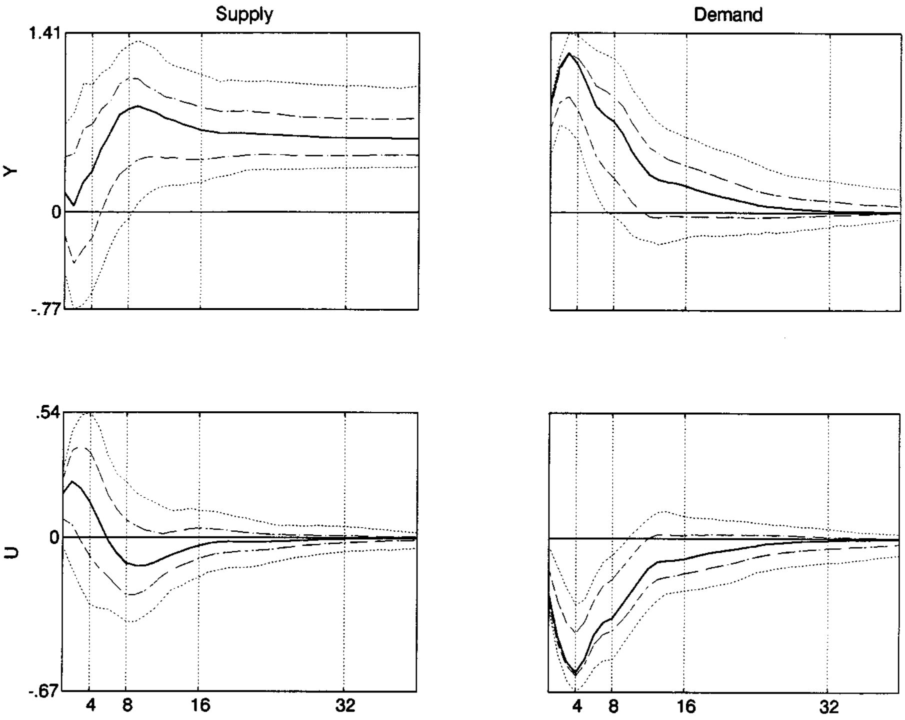

FIGURE 8.—68% and 95% other-percentile bootstrap bands, B-Q Model.

TABLE VII

COMPARISON OF INTERVAL LENGTHS, BLANCHARD-QUAH MODEL

	Ratios of Mean Lengths: Bayesian/Bootstrap			
t	*Y* to *S*	*U* to *S*	*Y* to *D*	*U* to *D*
1	1.009	1.004	.996	1.008
2	1.005	1.001	.985	1.006
3	.997	.999	.979	.993
4	.990	.992	.979	.984
6	.972	.981	.997	.984
8	.962	.975	.991	.994
12	.972	.985	.995	.999
16	.979	.983	1.005	1.016
24	.976	.942	.984	1.001
32	.963	.886	.958	.969

Note: Monte Carlo standard errors of these figures vary, but none are over .06.

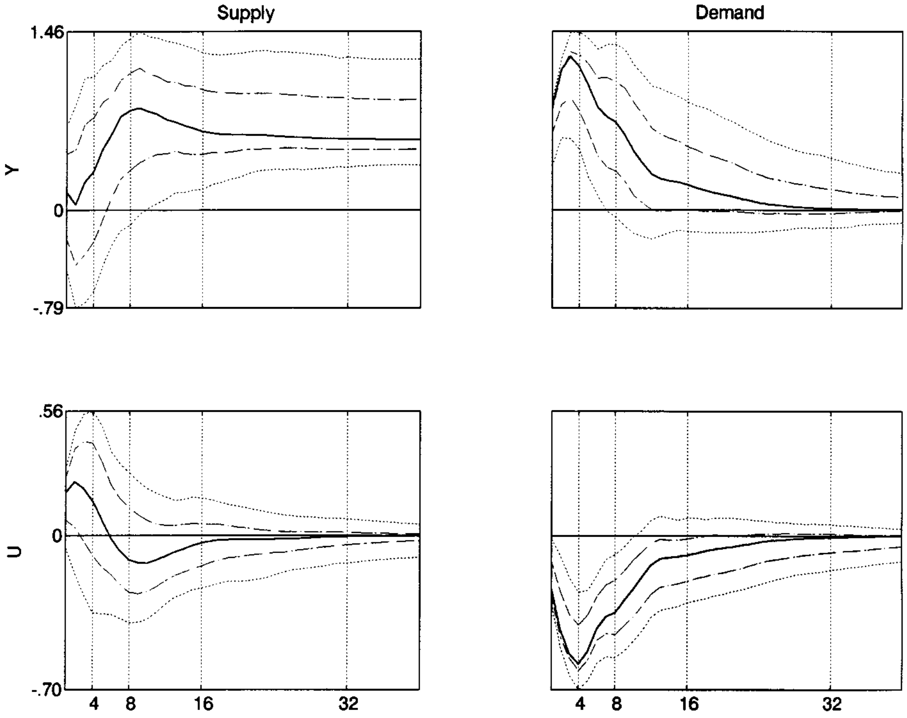

FIGURE 9.—68% and 95% bias-corrected bootstrap bands, B-Q Model.

TABLE VIII

BAYESIAN AND BOOTSTRAP INTERVALS AS CONFIDENCE REGIONS, BLANCHARD-QUAH MODEL

	Bootstrap Interval Coverage Probabilities					Bayesian Interval Coverage Probabilities			
t	*Y* to *S*	*U* to *S*	*Y* to *D*	*U* to *D*	*t*	*Y* to *S*	*U* to *S*	*Y* to *D*	*U* to *D*
1	.657	.650	.050	.632	1	.653	.662	.273	.672
2	.648	.653	.108	.425	2	.658	.655	.270	.618
3	.647	.642	.225	.295	3	.650	.668	.372	.462
4	.657	.642	.337	.297	4	.658	.660	.445	.448
6	.678	.647	.550	.452	6	.642	.637	.615	.505
8	.670	.672	.593	.525	8	.647	.623	.652	.600
12	.687	.705	.630	.635	12	.690	.705	.658	.632
16	.728	.775	.645	.707	16	.752	.792	.693	.725
24	.747	.845	.783	.780	24	.768	.860	.845	.853
32	.737	.910	.928	.865	32	.760	.923	.952	.920

Note: Monte Carlo standard error of a .68 frequency with 600 draws is .019. Bootstrap intervals have nominal coverage probability of .68. Bayesian intervals are flat-prior, equal-tail, .68 posterior probability intervals.

shown in Table IX. Bands on the first component are shown in Figure 10. These bands have the same general shape as those in Figure 7, but are for the most part narrower, reflecting the fact that they describe only one, not completely dominant, component of variation. It is interesting that the *Y*-to-supply response has a wider band on this first component for large *t* than one sees in the pointwise bands of Figure 7. This can only occur because of non-Gaussianity in the posterior distribution.

From Figure 11, showing the second component of variation, we see that uncertainty about shape of responses in this model is quite high, particularly for the response of *Y* to supply. One might get the impression from the pointwise bands that the main uncertainty is about the level, not the shape of the

TABLE IX

VARIANCE DECOMPOSITIONS FOR BQ MODEL RESPONSES

0.7147	0.2029	0.4092	*Y* to supply
0.7487	0.1319	0.0654	*U* to supply
0.6165	0.2002	0.0984	*Y* to demand
0.4555	0.2117	0.1835	*U* to demand

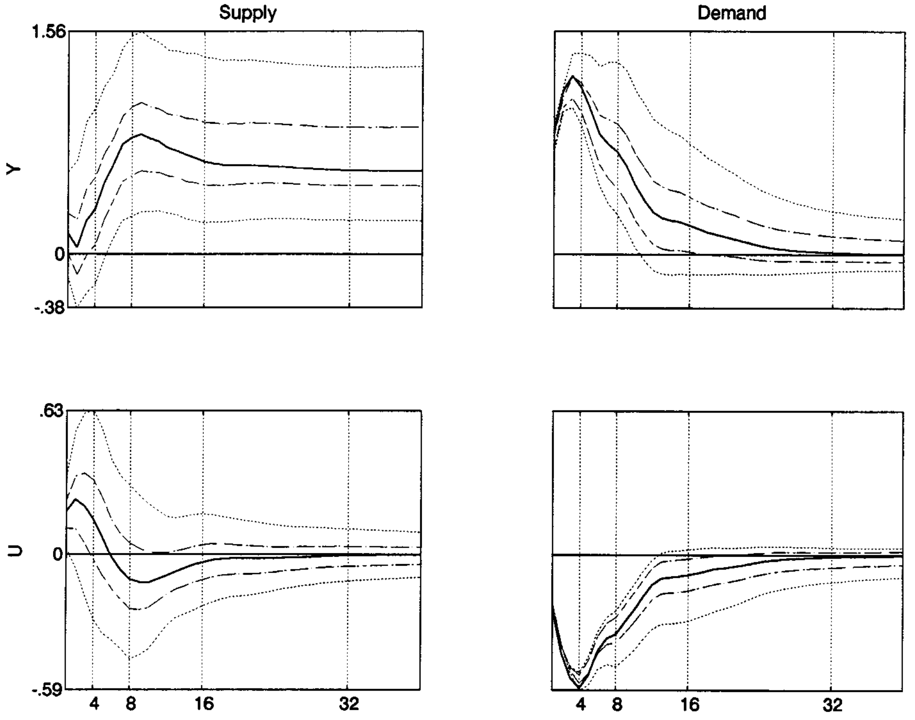

FIGURE 10.—First component: .68 and .95 probability bands, B-Q Model.

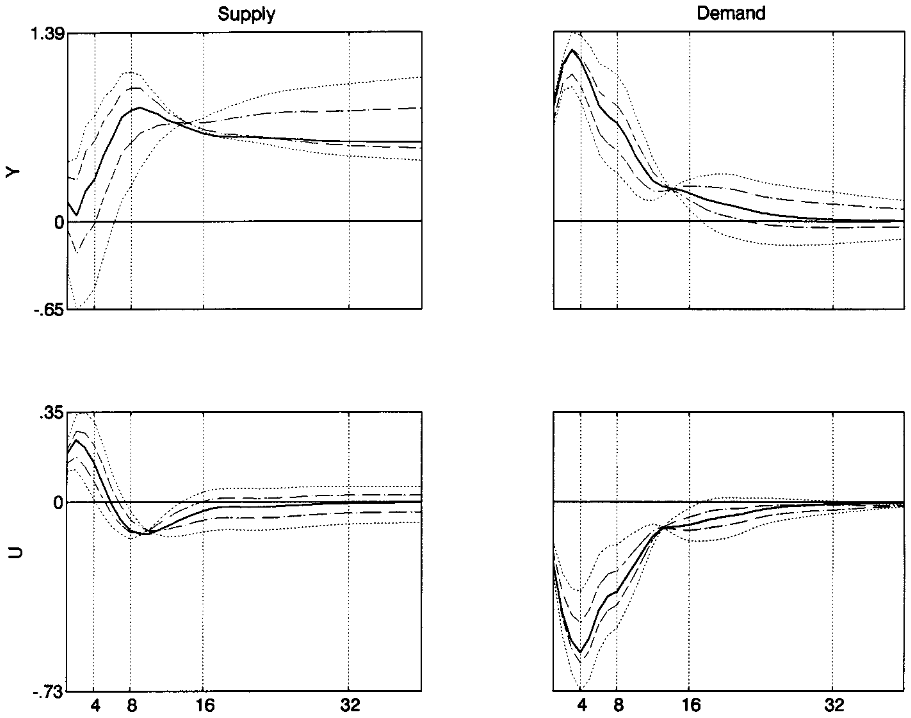

FIGURE 11.—Second component: .68 and .95 probability bands, B-Q Model.

Y-to-supply response, with more uncertainty about the initial than the later levels.

Figure 11 makes it clear that the very negative initial responses that are quite likely are associated with stronger positive long run responses. We display the third component of variation in these responses in Figure 12. These components account for 5–20% of the traces of the covariance matrices, according to the third column of Table IX. The three component plots together show that the shape of these responses is quite uncertain. Despite the widths of the bands being narrower than those in the Y-$M1$ mode, and despite the similar degree of smoothness in the plotted bands, the data are seen through the decomposition graphs to contain much sharper information about the shape of the responses in the Y-$M1$ model than do the data for the B-Q model. We might have suspected this difference from the fact that the B-Q model is fit to data on *Y growth*, while the Y-$M1$ model uses *Y levels*. This makes the Y-$M1$ data smoother, implying its impulse responses should be smoother. However, in a multivariate model it is dangerous to assume that impulse responses always inherit the degree of smoothness we see in plotted data. And there was no way to see, from the plotted pointwise responses alone, that there was this kind of difference in the information the data carry about the impulse responses of the two models.

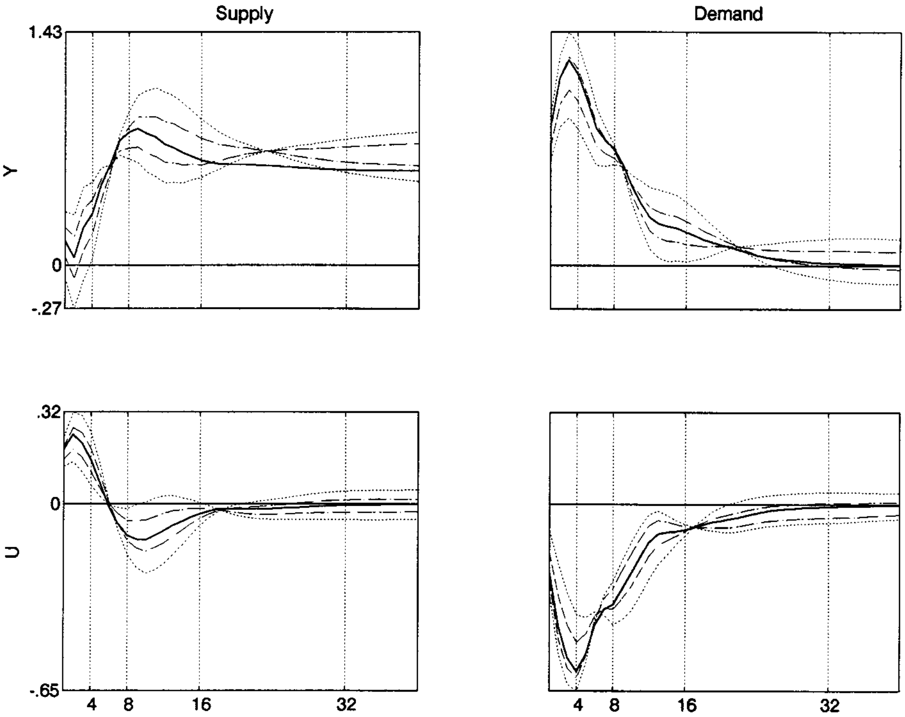

FIGURE 12.—Third component: .68 and .95 probability bands, B-Q Model.

With these simple models we hope to have made our case that likelihood-characterizing error bands constructed as flat-prior posterior probability intervals are feasible and can be a considerable improvement on bootstrap-based confidence intervals as descriptors of likelihood shape. We also hope to have demonstrated that information about shape beyond that in pointwise bands is useful and not reliably obtained by methods based on eyeball and intuition.

We now turn to considering a model closer to the scale now in use in analyzing macroeconomic policy. For this larger model we have overidentifying restrictions on parameters. It turns out that this makes computing error bands more demanding, in ways that have not been recognized in some of the existing literature.

8. MONTE CARLO METHODS FOR POSTERIOR PROBABILITIES IN OVERIDENTIFIED MODELS

A. Theory

We consider linear simultaneous equations models of the form

$$\Gamma(L)y(t) = \varepsilon(t). \tag{10}$$

We take

$$\varepsilon(t) | y(s), s < t \sim N(0, \Lambda), \tag{11}$$

with Λ diagonal. We assume Γ_0 to be nonsingular, so that (10) provides a complete description of the conditional distribution of $y(t)$ given $y(s)$, $s < t$ and can be solved by multiplying through on the left by Γ_0^{-1} to produce the reduced form

$$B(L)y(t) = u(t), \tag{12}$$

in which $B_0 = I$ and $u(t)$, while still uncorrelated with past y's, has a covariance matrix that is not in general diagonal, being given by

$$\Sigma = \Gamma_0^{-1} \Lambda \Gamma_0'^{-1}. \tag{13}$$

We assume the system is a finite-order autoregression, meaning that there is a $k < \infty$ such that $\Gamma_j = B_j = 0$ for all $j > k$.

The p.d.f. for the data $y(1), \ldots, y(T)$, conditional on the initial observations $y(-k+1), \ldots, y(0)$, is proportional to q as defined by

$$q(B, \Sigma) = |\Sigma|^{-T/2} \exp\left[-\tfrac{1}{2} \operatorname{trace}(S(B)\Sigma^{-1}) \right], \tag{14}$$

$$\hat{u}(t; B) = B(L)y(t), \tag{15}$$

$$S(B) = \sum_{t=1}^{T} \hat{u}(t, B)\hat{u}(t, B)'. \tag{16}$$

For a given sample, (14) treated as a function of the parameters B and Σ is the likelihood function. Its form is exactly that of the likelihood for a regression with Gaussian disturbances and strictly exogenous regressors, a classic model for which Bayesian calculations are well-discussed in the literature.[20] The RATS program includes routines to implement Monte Carlo drawing from the joint distribution of B and Σ and use of those draws to generate a Monte Carlo sample from the posterior distribution of impulse responses.[21]

The impulse responses for the model, defined by (5) above, are in this case the coefficients of

$$B^{-1}(L)\Gamma_0^{-1}\Lambda^{\frac{1}{2}}, \tag{17}$$

[20] See, e.g., Box and Tiao (1973, Chapter 8) for the theory.

[21] Box and Tiao (1973) recommend using a Jeffreys prior on Σ, which turns out to be proportional to $|\Sigma|^{-(m+1)/2}$. The packaged RATS procedure uses instead $|\Sigma|^{-(m+\upsilon+1)/2}$, where υ is the number of estimated coefficients per equation. Phillips (1992) suggests using the joint Jeffreys prior on B and Σ, which in time series models (unlike models with exogenous regressors) is not flat in B. The Phillips suggestion has the drawback that the joint Jeffreys prior is computationally inconvenient and changes drastically with sample size, making it difficult for readers to compare results across data sets. We therefore prefer the Box and Tiao suggestion in principle, though they point out (p. 44) that even in models with exogenous regressors mechanical use of Jeffreys priors can lead to anomalies. In this paper, to keep our results as comparable as possible to the existing applied literature, we have followed the RATS procedure's choice of prior.

where the $\Lambda^{\frac{1}{2}}$ factor scales the structural disturbances to have unit variance, or equivalently converts the responses so they have the scale of a response to a disturbance of "typical" (one-standard-deviation) size. Equation (13) gives us a relation among Σ, Γ, and Λ. Because Σ is symmetric, Γ_0 and Λ have more unrestricted coefficients than Σ. An exactly identified VAR model is one in which we have just enough restrictions available to make (13) a one-one mapping from Σ to Γ_0 and Λ. In this case, sampling from the impulse responses defined by (17) is straightforward: sample from the joint distribution of B and Σ by standard methods, then use the mapping defined by (13) and the restrictions to convert these draws from the distribution of impulse responses. The most common use of this procedure restricts Γ_0 to be triangular, solving for $\Gamma_0^{-1}\Lambda^{\frac{1}{2}}$ by taking a Choleski decomposition of Σ.

When the model is not exactly identified, however, reliance on the standard methods and programs that generate draws from the joint distribution of the reduced form parameters is no longer possible. A procedure with no small-sample rationale that does use the standard methods has occurred independently to a number of researchers (including ourselves) and been used in at least two published papers (Gordon and Leeper (1994), Canova (1991)). We will call it the naive Bayesian procedure. Because the method has a misleading intuitive appeal and may sometimes be easier to implement than the correct method we describe below, we begin by describing it and explaining why it produces neither a Bayesian posterior nor a classical sampling distribution.

In an overidentified model, (13) restricts the behavior of the true reduced-form innovation variance matrix Σ. It remains true, though, that the OLS estimates $\hat{B}$ and $\hat{\Sigma}$ are sufficient statistics, meaning that the likelihood depends on the data only through them. Thus maximum likelihood estimation of B, Γ_0, and Λ implies an algorithm for mapping reduced form $(\hat{B}, \hat{\Sigma})$ estimates into structural estimates $(B^*, \Gamma_0^*, \Lambda^*)$ that satisfy the restrictions. Often there are no restrictions that affect B, so that $\hat{B} = B^*$. The naive Bayesian method proceeds by drawing from the unrestricted reduced form's posterior p.d.f. for (B, Σ), then mapping these draws into values of (B, Γ_0, Λ) via the maximum likelihood procedure, as if the parameter values drawn from the unrestricted posterior on (B, Σ) were instead reduced from parameter estimates. The resulting distribution is of course concentrated on the part of the parameter space satisfying the restrictions, but is not a parametric bootstrap classical distribution for the parameter estimates, because the posterior distribution for (B, Σ) is not a sampling distribution for $(\hat{B}, \hat{\Sigma})$. It is not a true posterior distribution because the unrestricted posterior distribution for (B, Σ) is not the restricted posterior distribution, and mapping it into the restricted parameter space via the estimation procedure does not convert it into a restricted posterior distribution.

The procedure does have the same sort of asymptotic justification that makes nearly all bootstrap and Bayesian methods of generating error bands asymptotically equivalent from a classical point of view for stationary models, and it is probably asymptotically justified from a Bayesian viewpoint as a normal approximation even for nonstationary models. To see this, consider a simple normal linear estimation problem, where we have a true parameter β, an unrestricted estimate distributed as $N(\beta, \Omega)$, and a restriction $R\beta = \gamma$ with R $k \times m$. The restricted maximum likelihood estimate is then the projection on the $R\beta = \gamma$ manifold of the unrestricted ML estimate $\hat{\beta}$, under the metric defined by Ω, i.e.

$$\hat{\beta}^* = \Phi(\Phi'\Omega^{-1}\Phi)^{-1}\Phi'\Omega^{-1}\hat{\beta} + M(M'\Omega^{-1}M)^{-1}\gamma, \tag{18}$$

where $M = \Omega R'$ and Φ is chosen to be of full column rank $m - k$ and to satisfy $R\Phi = 0$. The sampling distribution of $\hat{\beta}^*$ is then in turn normal, since it is a linear transformation of the normal $\hat{\beta}$. In this symmetrically distributed, pure location-shift problem, the unrestricted posterior on β has the same normal p.d.f., centered at $\hat{\beta}$, as the sampling p.d.f. of $\hat{\beta}$ about β. We could make Monte Carlo draws from the sampling distribution of $\hat{\beta}^*$ by drawing from the sampling distribution of $\hat{\beta}$, the unrestricted estimate, and projecting these unrestricted estimates on the restricted parameter space using the formula (18). But since in this case the posterior distribution of $\hat{\beta}$ and its sampling distribution are the same, drawing from the posterior distribution in the first step would give the same correct result. And since in this case the restricted posterior has the same normal shape about $\hat{\beta}^*$ that the sampling distribution of $\hat{\beta}^*$ has about β, the simulated distribution matches the posterior as well as the sampling distribution of the restricted estimate.

The naive Bayesian method for sampling from the distribution of impulse responses rests on confusing sampling distributions with posterior distributions, but in the case of the preceding paragraph this would cause no harm, because the two kinds of distribution have the same shape. For stationary models, distribution theory for $\hat{\Sigma}$ and $\hat{B}$ is asymptotically normal, and differentiable restrictions will behave asymptotically as if they were linear. So the case considered in the previous paragraph becomes a good approximation in large samples. For stationary or nonstationary models, the posterior on Σ is asymptotically normal, so the naive Bayesian method is asymptotically justified from a Bayesian point of view.

But in this paper we are focusing on methods that produce error bands whose possible asymmetries are justifiably interpreted as informative about asymmetry in the posterior distribution of the impulse responses. Asymmetries that appear in bands generated by the naive Bayesian method may only turn out to be evidence that the asymptotic approximations that might justify the method are not holding in the sample at hand.

It is important to note that, though the naive Bayesian method will eventually work well in large enough samples, meaning in practice situations where the sample determines estimates precisely, it can give arbitrarily bad results in particular small-sample situations. For example, in a 3-variable model that is exactly identified by the order condition, via three zero restrictions on Γ_0 in the pattern shown in Table X (where x's indicate unconstrained coefficients) it is known that the concentrated or marginalized likelihood function, as a function

TABLE X

EXAMPLE IDENTIFICATION

1	0	x
x	1	0
0	x	1

Note: x's are unconstrained coefficients.

of Γ_0, Λ, generically has two peaks of equal height. (See Bekker and Pollock (1986) and Waggoner and Zha (1998).) This pattern is perhaps well enough known that if an applied three-variable model took this form, the model would be recognized as globally unidentified. But the fact that it is globally unidentified does not mean that the data are uninformative about the model. A correct sampling from the likelihood or posterior p.d.f., using the method we have proposed will combine information from both peaks. If the peaks are close together in impulse response space, their multiplicity may not be a problem. If they are far apart, the averaged impulse responses will have wide error bands, correctly traced out by integrating the likelihood. The naive Bayesian algorithm, however, would simply pick one of the peaks for each draw arbitrarily, according to which peak the equation-solving algorithm converged to. This would easily turn out to be the same peak every, or nearly every, time, so that uncertainty would be greatly underestimated.[22]

This pattern of identifying restrictions also creates another difficulty for the naive Bayesian procedure. Even though there are as many unconstrained coefficients as distinct elements in Σ, so that the order condition for identification is satisfied, Γ_0 matrices of this form do not trace out the whole space of positive definite Σ's. That is, there are Σ's for which (13) has no solution subject to these constraints. In drawing from the unconstrained posterior distribution of Σ, the naive Bayesian procedure would occasionally produce one of these Σ's for which there is no solution. If the usual equation-solving approach for matching Γ_0 and Λ to Σ in a just-identified model is applied here, it will fail.

We raise this simple 3×3 case just to show that models that create problems for the naive Bayesian method exist. More generally, models in which likelihoods have multiple peaks do arise in overidentified models, and they create difficulties for the naive Bayesian approach. The difficulties are both numerical —in repeatedly maximizing likelihood over thousands of draws it is impractical to monitor carefully which peak the algorithm is converging to—and analytical

[22] While this example may appear special, it can easily be embedded in a larger model. If a larger model had a 3×3 block with this pattern of zero restrictions in its upper left corner, but was not block triangular, it could be globally identified. Nonetheless, if the true coefficient matrix were close to block diagonality, there would most likely be two peaks in the likelihood, of somewhat different, but similar height. This would create the same possibility as in the simple 3×3 example for the naive Bayesian procedure to reflect the likelihood shape only near one of the peaks.

—when there are multiple peaks the asymptotic approximations that can justify the naive Bayesian procedure are clearly not accurate in the current sample.

To describe a correct procedure for generating Monte Carlo draws from the Bayesian posterior for the parameters in (10), we begin by introducing a reparameterization. In place of (10) we use

$$A(L)y(t) = \eta(t), \tag{19}$$

where $A = \Lambda^{-\frac{1}{2}}\Gamma$ and $\eta(t) = \Lambda^{-\frac{1}{2}}\varepsilon(t)$ so that $\operatorname{var}(\eta(t)) = I$. There is no real reduction in the number of free parameters, because the diagonal of Γ_0 is always normalized to a vector of ones so that an unrestricted A_0 has the same number of parameters as a diagonal Λ together with a normalized Γ_0. There are a number of reasons to prefer this parameterization. It simplifies the mapping (17) between reduced form and structural parameters. The usual parameterization can give a misleading impression of imprecision in the estimate of a row of Γ_0 if the normalization happens to set to 1 a particularly ill-determined coefficient in the corresponding row of A_0. But the main reason for introducing this parameterization here is that the likelihood is not in general integrable when written as a function of (B, Γ_0, Λ), requiring some adjustment of the flat prior to allow formation of a posterior distribution, whereas it is integrable as a function of (B, A_0).[23]

We can rewrite the likelihood (14) as

$$|A_0|^T \exp\left\{-\tfrac{1}{2}\operatorname{trace}\left(A_0' A_0 S(\hat{B}\right) - \tfrac{1}{2}\operatorname{trace}\left((B-\hat{B})' X'X(B-\hat{B}) A_0' A_0\right)\right\}. \tag{20}$$

Taking the prior as flat in B and A_0, we can integrate over B to obtain the marginal posterior on A_0,

$$p(A_0) \propto |A_0|^{T-v} \exp\left[-\tfrac{1}{2}\operatorname{trace}\left(A_0 S(\hat{B}) A_0'\right)\right]. \tag{21}$$

Here as with the reduced-form model we have followed the widely used RATS code in dropping the "degrees of freedom correction" $-v$ in (21). This can be thought of as in effect using $|A_0|^v$ as an improper prior, or as the consequence of starting with a flat prior on the coefficients of $A(L)$, then converting to a parameterization in terms of A_0 and $B(L)$. As can be seen by comparing (20) with (21), this has the effect of making the marginal posterior on A_0 proportional to the concentrated likelihood and thereby eliminating possible discrepancies between posterior modes and maximum likelihood estimates.

[23] In general, the likelihood, with B and Λ integrated out, is $O(1)$ in individual elements of Γ_0 at all sample sizes. This does not mean that it fails to be asymptotically normal—the likelihood does become small outside a region around the true parameter values. There are also special cases where the likelihood is integrable, e.g. where Γ_0 is normalized with ones down the diagonal and restricted to be triangular. For details, see the earlier draft of this paper, Cowles Foundation Discussion Paper No. 1085. Of course, we could use the standard parameterization and retain integrability if we made the prior flat in A_0, then transformed back to the (Γ_0, Λ) parameter space, including the appropriate Jacobian term $|\Lambda|^{-(m+1)/2}$. But the easiest way to explain the appearance of this term would be to derive it from the A_0 parameterization, which is in any case more well-behaved.

Expression (21) is not in the form of any standard p.d.f. To generate Monte Carlo samples from it, we first take a second-order Taylor expansion of it about its peak, which produces the usual Gaussian approximation to the asymptotic distribution of the elements of A_0. Because this is not the true form of the posterior p.d.f, we cannot use it directly to produce our Monte Carlo sample. One approach is importance sampling, in which we draw from the Gaussian approximation, or from a multivariate t with the same covariance matrix, but weight the draws by the ratio of (21) to the p.d.f. from which we draw. The weighted sample c.d.f. then approximates the c.d.f. corresponding to (21).[24] The weights in practice vary rather widely, so that a given degree of Monte Carlo sampling error in impulse response bands computed this way generally requires many times as many Monte Carlo draws as for a reduced form model where weighting is not required. We found this straightforward importance sampling procedure not to work well for our task.

We have instead used a version of the random walk Metropolis algorithm for Markov chain Monte Carlo (MMCMC) sampling from a given p.d.f. Details of the method are set out in Waggoner and Zha (1997).[25] The simulation results we present ought to be, by several measures, very accurate, but they did occupy substantial computing time (11 hours on a Pentium II 233 MHz laptop). We expect that further refinements in the sampling algorithm can reduce this time.

Note that it is also possible to compute the error bands without any weighting. This is yet another example of a method for computing error bands that is asymptotically justified, but admits no rationale for interpreting its asymmetries as providing information about small-sample deviations from normality. It is even more likely than the naive Bayesian procedure to produce unrealistically tight bands in the presence of multiple local maxima.

Though in switching to the A_0 parameterization we have eliminated the need to choose a normalization in the usual sense, there is still a discrete redundancy in the parameterization: the sign of a row of A_0 can be reversed without changing the likelihood function. It may seem that a choice of normalization

[24] This idea, importance sampling, has seen wide use since its introduction into econometrics by Kloek and Van Dijk (1978).

[25] The algorithm used three separate randomly drawn starting points and a multivariate t distribution with 6 degrees of freedom for the jump distribution on changes in A_0. Because the distribution of the remaining parameters conditional on A_0 is Gaussian, we needed the MMCMC algorithm only for the A_0 parameters. The covariance matrix of the t distribution was set at the Hessian of the log likelihood at the peak of the log likelhood, scaled by $-.25$. We used 480,000 draws for each starting point, discarding the first half of each run, so that final results are based on 720,000 draws. The "potential reduction scale" of Gelman, Carlin, Stern and Rubin (1995) is between 1 and 1.0085 for each of the 20 free parameters in A_0, and for over half of the parameters it is between 1 and 1.001. The frequency of accepted jumps was about .49 in each of the three chains. A measure of effective Monte Carlo sample size, which provides some guidance as to the reliability of the tabulated quantiles, is the Monte Carlo sample size times the ratio of "within chain" to "between chain" estimates of the variance of a parameter. For our total sample size of 720,000, the measure ranges from 176, for the element of A_0 with the worst value of the potential reduction scale, to 772 at the average value of the potential reduction scale across all parameters.

cannot have much effect on substantive results, but this is not true. In the conventional parameterization, with diagonal elements of Γ_0 normalized to 1, it is widely understood that an equation with statistically well-determined coefficients on most of its variables can be converted to one in which all coefficients seem ill-determined by normalizing the equation on a variable whose coefficient is insignificantly different from zero. This is because such a normalization makes the sign of all the coefficients in the equation nearly indeterminate. A similar phenomenon arises with impulse responses. Casual choice of normalization can lead to estimates that all responses to, say, a policy shock are ''insignificant,'' when a better normalization would make it clear that the responses are actually sharply determined. Here we follow Waggoner and Zha (1997) in choosing a normalization for each draw that minimizes the distance of A_0 from the ML estimate of A_0. Our experience confirms their suggestion that this method will tend to hold down spurious sign-switching of impulse responses and thereby deliver sharper results.[26]

B. *Results for a Six-Variable Model*

An earlier paper by one of us (Sims (1986)) contains an example of an overidentified VAR model with an interesting interpretation. (It identifies a money supply shock that produces both a liquidity effect—an initial decline in interest rate—and a correctly signed price effect—inflation following a monetary expansion.) The paper contains two sets of identifying restrictions, and we show calculations for the identification labeled version 2 in the original paper.

Figure 13 shows impulse responses for this model, with .68 and .95 flat-prior probability bands. The first column shows that the identified monetary policy shocks have an expected pattern of effects, raising interest rates at least initially, lowering the money stock, lowering output for at least a year or so, lowering the inflation rate, though perhaps not by much, raising the unemployment, and at least initially lowering investment. The second column was meant to be interpreted as the effect of a money demand shock. The wide and strongly asymmetric error bands show that the effects of this disturbance are ill-determined. The strong asymmetry suggests that possibly the likelihood has multiple peaks, with qualitatively different ''*MD*'' columns across the peaks. The Y and I columns display private-sector shocks that cause the Fed to raise interest rates. The Y column is a shock that raises output, but not prices, while the I shock column raises output and the inflation rate and has a more sharply determined, though possibly not stronger, interest rate response.

[26] The method we have used minimizes the Euclidean distance between the ML estimate and the normalized draw. This makes the method sensitive to the scale of the variables. All our variables have variances of about the same scale, so this is not in itself a problem, but it would be better systematically to allow for scale in the metric. This could be accomplished by, say, using the inverse Hessian of the log likelihood at the maximum to define the distance metric. When we used other, apparently reasonable, normalization rules, the effects on the estimated error bands were in every case clearly visible, usually in the direction of widening them.

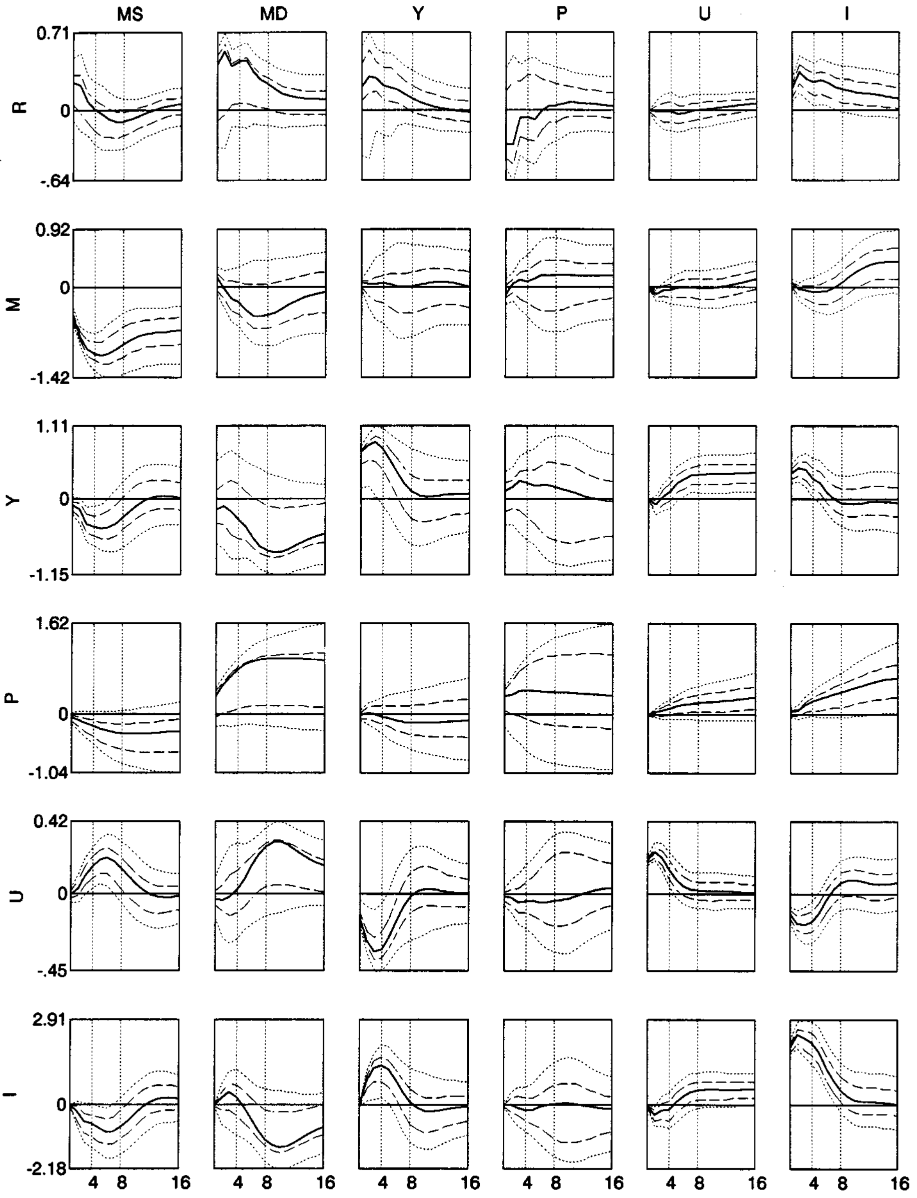

FIGURE 13.—Pointwise .68 and .95 probability bands, Six-Variable Model.

Decomposition of these responses into components provides useful insight. They are more like the Blanchard-Quah responses than the Y-$M1$ responses in that fairly low proportions of the traces of the covariance matrices for these responses are accounted for by the first few eigenvalues. The range is from .49 to .97, with a mean of .75. Since our focus here is not a detailed substantive assessment of this model, we omit display of decompositions for individual responses. But it is worthwhile to show the usefulness of cross-variable decompositions. We stacked the responses to the monetary policy shock of R, $M1$, P, and Y and decomposed them jointly. For this decomposition the first six eigenvalues account for the following proportions of the variance matrix trace: 0.52, 0.25, 0.06, 0.05, 0.05, 0.02. Figure 14 shows .95 probability bands for these six components.

The most interesting aspect of this graph emerges only with careful attention to the different dotted line types.[27] For the first component, in column 1, the top lines in plots of the responses of M and Y match the bottom lines in the plots of responses of R and P. Thus more positive responses of R and P are associated with more negative responses of M and Y. The first three variables are consistent in suggesting that the main component of uncertainty is simply the uncertainty about the strength of the effect of the shock—is the increase in R, decline in M, and initial decline in Y large, or small? But the P component shows the opposite pattern—large R, M, and Y responses are associated with weak P responses, and vice versa. This pattern could be relevant to interpretation of the model's results, and would be very important to include in characterizations of uncertainty about conditional forecasts from the model used in policy analysis.

The second column weights most heavily on M and P, indeed is more important for M than is the first component. Here M and P move together. This component suggests that there is a substantial uncertainty about the general effect on nominal variables of monetary contraction, independent of uncertainty about how strongly monetary contraction affects the interest rate and output. The third component affects mostly the output response, the fourth mostly the M response, and the fifth mostly the R response, but all are modest in size.

This model shows only moderate simultaneity, so that difference between these correctly computed bands and bands computed by the naive Bayesian method or by using unweighted draws from the asymptotic normal approximation to the p.d.f. for A_0 are correspondingly modest. The bands around responses to monetary policy disturbances are particularly similar across methods, probably because these responses are relatively sharply determined. Differences are more clearly visible in the responses to MD and Y shocks, which are more ill-determined. In Figure 15 we compare bands constructed correctly by Metropolis chain, by the naive Bayesian method, and by unweighted draws from the Gaussian approximation based on the Hessian of the log likelihood at the MLE. One response that is of interpretive interest for which bands come out differently across these methods is that of R to a Y shock. The Y shock is major source of variation in output, and a positive response of R to this shock suggests

[27] Our own working versions of all these graphs use color to allow a more detailed distinction among the displayed lines. These color graphs are available, together with earlier, quite different and not completely obsolete, versions of the entire paper, at http://www.princeton.edu/~sims.

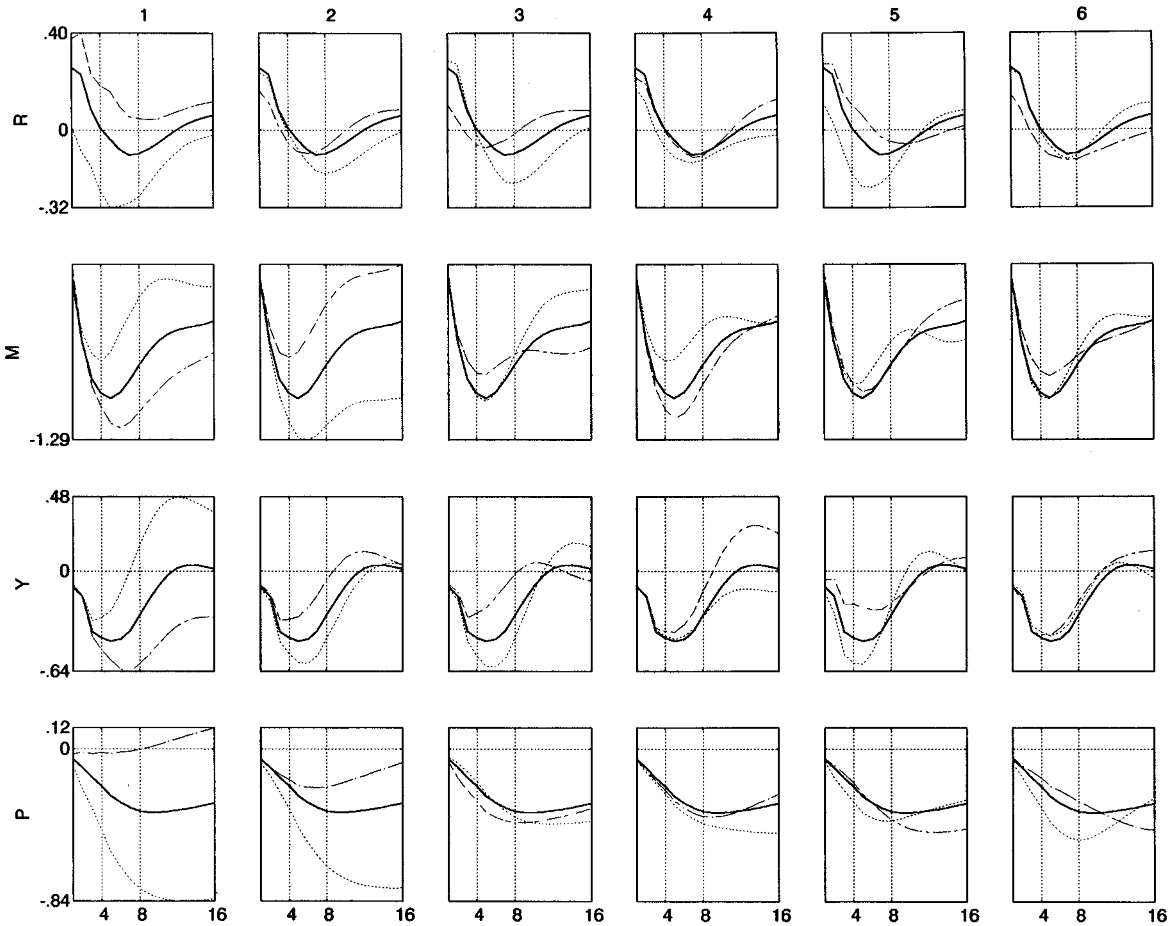

FIGURE 14.—Components 1–6 of joint response of R, M, P, Y to monetary policy, .95 bands, Six-Variable Model.

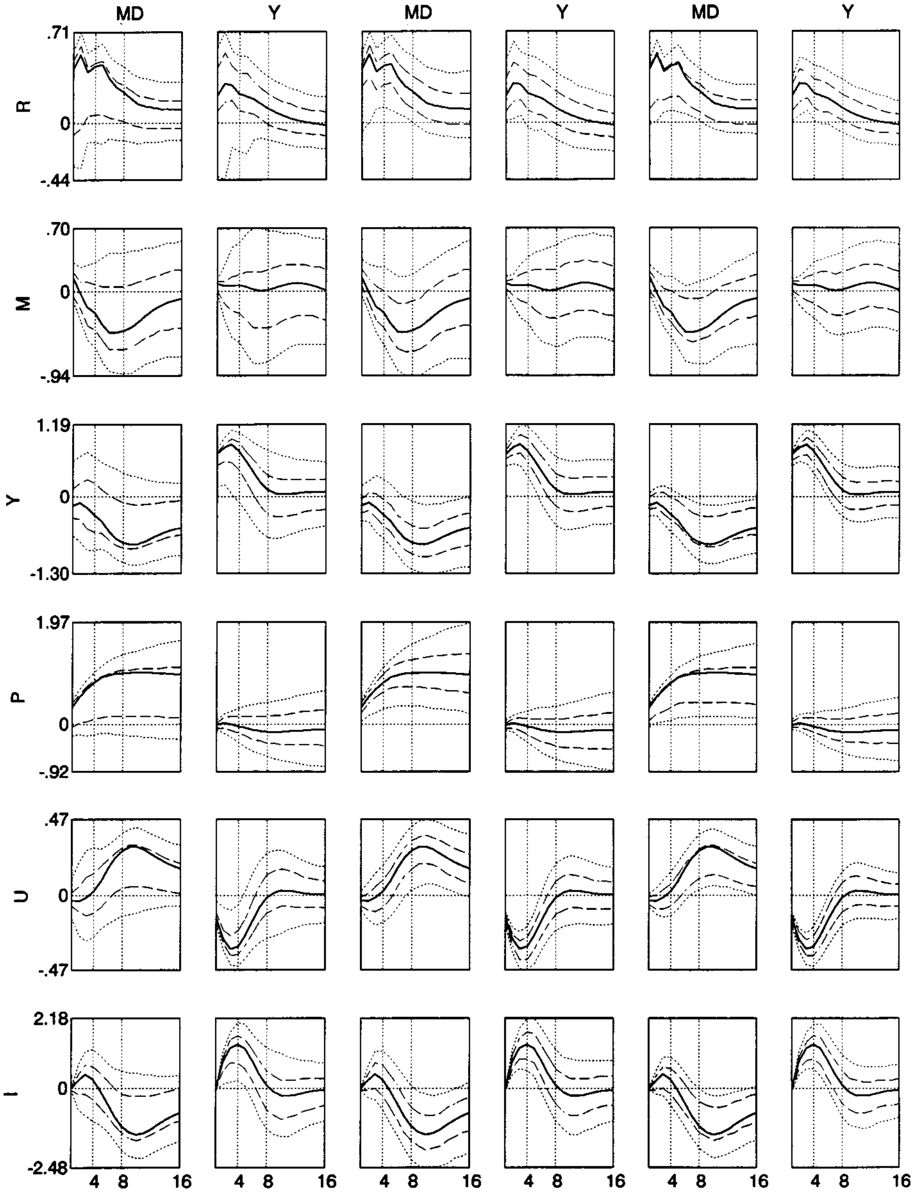

FIGURE 15.—Six-Variable Model, responses to money demand and *Y*, comparing correct, naive, and unweighted Bayesian bands.

that monetary policy tends to tighten when output expands. But note that in Figure 13 the 95% band extends well into negative range for the initial *R*-to-*Y* response, indicating considerable uncertainty about the strength of this response. The two other methods, by constrast, show 95% bands that in each case are entirely positive. The responses to *MD* are of less substantive interest, since the results seem to show that this set of responses is so ill-determined that interpretation is dubious. However, it can be seen that there is a general tendency for the third and fifth columns, using the naive and unweighted methods, to give narrower bands than the first column. Finally, the band for the response of price to MD in the third column, corresponding to the naive Bayesian procedure, shows hardly any skewness, while the likelihood in fact shows very strong skewness, with the 68% band lying almost entirely below the point estimate.[28]

We also checked bootstrap methods for this model. The other-percentile method without bias correction does surprisingly well. Where the likelihood shows skewness, the uncorrected other-percentile bands tend to exaggerate the skewness, probably because of their bias-amplification tendency. These bands show the point estimate almost entirely outside a 68% band in five instances and almost entirely outside a 95% band (the skewed P to MD response) in one instance, but the qualitative picture presented by these bands is otherwise in line with the likelihood-based bands. The other-percentile method with bias correction performs badly. In a relatively large model like this, there are likely to be a few roots at or just outside the boundary of the region of stationarity. Bias correction of the bootstrap distribution of $A(L)$ tends to push these roots unrealistically far into the nonstationary region. The result is that bands from the bias-corrected bootstrap tend to splay out rapidly with increasing t. For small t, disagreement of these bands with likelihood-based bands is more modest, but still important. The bands tend to be wider than the likelihood-based bands even for small t, and for the substantively important response of P to MS, the bias-corrected bootstrap includes 0 in its 68% bands, whereas the likelihood-based method (and all the others, as well) show 0 as well above the 68% band. Conclusions about whether the model shows a ‘‘price puzzle’’ might therefore be distorted by use of bias-corrected bootstrap intervals.

9. CONCLUSION

We have explained the reasons for computing likelihood-characterizing error bands in dynamic models, shown how to go beyond the standard pointwise bands to display shape information, shown that the methods we propose are computationally feasible, and shown that older approaches can fall short in characterizing actual likelihoods, even though they are known to provide accurate approximations in large enough samples. We hope to have convinced applied researchers that likelihood-characterizing bands are the best approach to providing error bands in time series models. We hope to have convinced theoretical researchers that improving likelihood-based approaches is a more productive

[28] We are confident that the differences discussed here are not artifacts of Monte Carlo sampling error. 8000 independent draws were used for the naive Bayesian and unweighted Bayesian bands, allowing quite precise estimates of the positions of bands. For the Metropolis chain estimates in the first two columns, we compared results for the pooled Monte Carlo sample with those for the three independent chains separately. The differences across the bands from the three chains were barely detectable by eye, so every pattern we discuss in the text is sharply determined.

use of intellectual effort than further attempts to provide confidence bands with asymptotic justification.

Department of Economics, Princeton University, Fisher Hall, Princeton, NJ 08544-1021; sims@princeton.edu; http://www.princeton.edu/~sims

and

Federal Reserve Bank of Atlanta, 104 Marietta Street, Atlanta, GA 30303; tzha@mindspring.com

Manuscript received October, 1994; final revision received August, 1998.

REFERENCES

BEKKER, PAUL A., AND D. S. G. POLLOCK (1986): "Identification of Linear Stochastic Models with Covariance Restrictions," *Journal of Econometrics*, 31, 179–208.

BERGER, JAMES O., AND ROBERT L. WOLPERT (1988): *The Likelihood Principle* (*2nd Edition*), Hayward, California: Institute of Mathematical Statistics.

BLANCHARD, O. J., AND D. QUAH (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79, 655–673.

BOSE, A. (1988): "Edgeworth Correction by Bootstrap in Autoregressions," *Annals of Statistics*, 16, 1709–1722.

BOX, G. E. P., AND GEORGE TIAO (1973): *Bayesian Inference in Statistical Analysis*. Reading, Mass.: Addison-Wesley.

CANOVA, FABIO (1991): "The Sources of Financial Crisis: Pre- and Post-Fed Evidence," *International Economic Review*, 32, 689–713.

DUFOUR, J.-M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1387.

EFRON, BRADLEY, AND ROBERT J. TIBSHIRANI (1993): *An Introduction to the Bootstrap*. New York and London: Chapman and Hall.

FAUST, JON (1996): "Near Observational Equivalence and Theoretical Size Problems with Unit Root Tests," *Econometric Theory*, 12, 724–731.

FERGUSON, THOMAS S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. New York and London: Academic Press.

GEWEKE, JOHN (1995): "Bayesian Comparison of Econometric Models," processed, University of Minnesota.

GELMAN, ANDREW, JOHN B. CARLIN, HAL S. STERN, AND DONALD B. RUBIN (1995): *Bayesian Data Analysis*. London: Chapman and Hall.

GORDON, D. B., AND E. M. LEEPER (1994): "The Dynamic Impacts of Monetary Policy: An Exercise in Tentative Identification," *Journal of Political Economy*, 102, 1228–1248.

HALL, PETER (1992): *The Bootstrap and Edgeworth Expansion*. New York, Berlin, Heidelberg: Springer-Verlag.

HILDRETH, CLIFFORD (1963): "Bayesian Statisticians and Remote Clients," *Econometrica*, 31, 422–439.

HOROWITZ, JOEL L., AND N. E. SAVIN (1998): "Empirically Relevant Critical Values for Hypothesis Tests: The Bootstrap To The Rescue," processed, Department of Economics, University of Iowa.

KILIAN, LUTZ (1998a): "Small-Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics*, 80, 186–201.

——— (1998b): "Pitfalls in Constructing Bootstrap Confidence Intervals for Asymptotically Pivotal Statistics," processed, Department of Michigan.

KLOEK, T., AND H. K. VAN DIJK (1978): "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, 46, 1–19.

KOOP, G. (1992): "Aggregate Shocks and Macroeconomic Fluctuations: A Bayesian Approach," *Journal of Applied Econometrics*, 7, 395–411.

KWAN, YUM K. (1998): "Asymptotic Bayesian Analysis Based on a Limited Information Estimator," City University of Hong Kong Discussion Paper, forthcoming in *Journal of Econometrics*.

LASTRAPES, W. D., AND G. SELGIN (1994): "Buffer-Stock Money: Interpreting Short-Run Dynamics Using Long-Run Restrictions," *Journal of Money, Credit and Banking*, 26, 34–54.

LOÈVE, MICHEL (1988): "Comment," in *The Likelihood Principle* (2nd Edition), by J. Berger and R. Wolpert. Hayward, CA: Institute of Mathematical Statistics.

LUTKEPOHL, H. (1990): "Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models," *The Review of Economics and Statistics*, 72, 53–78.

MITTNIK, S., AND P. A. ZADROZNY (1993): "Asymptotic Distributions of Impulse Responses, Step Responses, and Variance Decompositions of Estimated Linear Dynamic Models," *Econometrica*, 20, 832–854.

PHILLIPS, P. C. B. (1991): "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends," *Journal of Applied Econometrics*, 6, 333–364.

POTERBA, J. M., J. J. ROTEMBERG, AND L. H. SUMMERS (1986): "A Tax-Based Test for Nominal Rigidities," *American Economic Review*, 76, 659–675.

RATS, computer program available from Estima, 1800 Sherman Ave., Suite 612, Evanston IL 60201.

ROTHENBERG, THOMAS J., AND JAMES H. STOCK (1997): "Inference in a Nearly Integrated Autoregressive Model with Nonnormal Innovations," *Journal of Econometrics*, 80, 269–286.

RUNKLE, D. E. (1987): "Vector Autoregressions and Reality," *Journal of Business and Economic Statistics*, 5, 437–442.

SIMS, C. A. (1986): "Are Forecasting Models Usable for Policy Analysis?" *Federal Reserve Bank of Minneapolis Quarterly Review*, 10, 2–15.

WAGGONER, DANIEL F., AND TAO ZHA (1997): "Normalization, Probability Distribution, and Impulse Responses," Federal Reserve Bank of Atlanta Working Paper 97-11.

——— (1998): "Identification Issues in Vector Autoregressions," processed, Federal Reserve Bank of Atlanta.

WILKS, SAMUEL S. (1962): *Mathematical Statistics*. New York, London: John Wiley and Sons.

ZACKS, SHELEMYAHU (1971): *The Theory of Statistical Inference*. New York, London, Sydney, Toronto: John Wiley and Sons.