Volume Title: Business Cycles, Indicators and Forecasting

Volume Author/Editor: James H. Stock and Mark W. Watson, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-77488-0

Volume URL: http://www.nber.org/books/stoc93-1

Conference Date: May 3-4, 1991

Publication Date: January 1993

Chapter Title: A Nine-Variable Probabilistic Macroeconomic Forecasting Model

Chapter Author: Christopher A. Sims

Chapter URL: http://www.nber.org/chapters/c7192

Chapter pages in book: (p. 179 - 212)

4 A Nine-Variable Probabilistic Macroeconomic Forecasting Model

Christopher A. Sims

Beginning around 1980, Robert Litterman began forecasting aggregate macroeconomic variables using a small Bayesian vector autoregressive (BVAR) model. The model originally used six variables—the Treasury-bill rate, M1, the GNP deflator, real GNP, real business fixed investment, and unemployment. Litterman ceased forecasting with his model and turned the task over to me in 1986. At that time, his model had already changed, and I changed it further in 1987. This paper describes the current form of the model, explains why it changed as it did, and displays some measures of the model's performance since the major 1987 changes.[1]

The model differs in important respects from previous Bayesian VAR models that have been described in the literature (e.g., Litterman 1986, and Doan, Litterman, and Sims 1984). It accounts for nonnormality of forecast errors and allows for time-varying variances as well as time-varying autoregressive coefficients. According to its own likelihood function, it fits much better than the simpler earlier models. It implies much more time variation in autoregressive coefficients than the earlier models. Both within sample and out of sample it produces drastically better forecasts of the price level than the simpler models. For other variables, its advantages over the simpler models are smaller and uncertain.

Christopher A. Sims is professor of economics at Yale University and a research associate of the National Bureau of Economic Research.

This research was supported in part by the Institute for Empirical Macroeconomics at the Minneapolis Federal Reserve Bank and the University of Minnesota and also by National Science Foundation grant SES91–22355.

1. Joint work now under way by Richard Todd and me will provide a much more detailed assessment of the model's forecasting performance and of the contribution of its various components to its behavior.

4.1 A Brief History of These Models

Litterman's model performed remarkably well relative to forecasts prepared by commercial forecasting organizations using much more elaborate models (see Litterman 1986). In particular, as documented by McNees (1986), it performed better than commercial models for real GNP and unemployment but worse for the price level.

The model had not remained static in form from 1980 to 1986. Litterman had adapted the time-varying-parameters framework of Doan, Litterman, and Sims (1984) to his model. Also, it was easy to see in 1986 from graphs or tables of the forecasts that the model was extrapolating inflation at a long-run average rate, despite many quarters in a row of same-signed forecast errors. Thus, it was not surprising that McNees found other models doing better at forecasting inflation. Attempting to rectify this, Litterman added three variables to the original six—the trade-weighted value of the dollar, the Standard and Poors 500 stock price index, and a commodity price index.

With the model in this form, I took over preparing forecasts with it, starting in the fall of 1986. Litterman regularly evaluated his models by calculating measures of their forecast performance based on recursively updating their coefficients through the sample period, generating artificial "out-of-sample" forecasts. He had noted in these exercises a tendency for improvements in the retrospective forecast performance of the BVAR model for inflation to be accompanied by deterioration in its performance for real variables. He had chosen his additional variables aiming to minimize the real-variable deterioration while improving price forecasts. My own analysis suggested, however, that this attempt was not entirely successful. Furthermore, as I took over the model, it had been making a sequence of same-signed errors in forecasting real GNP, which, while not as serious as the earlier sequence of inflation errors, were disturbingly similar in pattern. I decided, therefore, to complicate the specification of the model in several ways, aiming to find a probability model that would track the shifts in trend inflation rates and productivity growth rates while still performing about as well for real variables as Litterman's original simple six-variable model.

The resulting model differs from Litterman's in several respects:

1. It allows for conditional heteroskedasticity (time-varying variances of disturbance terms).
2. It allows for nonnormality of disturbances. Specifically, it allows disturbances to be mixtures of two normal random variables.
3. It takes account of the connection of the constant term to the means of the explanatory variables using a "dummy initial observation," described below.
4. It uses the discrete-time process generated by time averaging of a continuous-time random walk as a prior mean, rather than using a discrete-time random walk.

5. Probably mainly as a result of the first three changes, it fits best with a great deal more implied time variation of parameters than Litterman found optimal with his model.

Likelihood is dramatically higher for this version of the model than for its predecessor. Simulated one- through eight-step-ahead forecasts from the sample period are about as good as or a bit better than with the previous model for real variables, much better for price variables, and slightly worse for interest rates.

4.2 Description of the Model

The data are a time series of $k \times 1$ vectors $X(t)$, determined by a state vector $\beta(t; i, j, s)$ and an equation disturbance $u(t; i)$ according to

$$(1) \quad X_i(t) = \sum_{j=1}^{k} \sum_{s=1}^{m} X_j(t - s)\beta(t; i, j, s) + \beta(t; i, j + 1, 1) + u(t; i).$$

I treat the β's and u's as stochastic processes that generate a distribution, conditional on initial X's, for the other observed X's. In principle, inference on all equations of the system should proceed jointly, as randomness in one equation could be correlated with randomness in other equations.[2] However, because it is computationally convenient, and because some tentative experiments have indicated little advantage from full-system estimation, estimation proceeds equation by equation. What I discuss below, therefore, although I call it the "likelihood," is usually the component of the likelihood corresponding to one equation under the assumption of independence across equations. Likelihood for the full system is then taken as the sum of these equation likelihoods and is the true full-system likelihood only under an assumption of independence of randomness across equations.

4.2 Form of the Distribution of Disturbances

Conditional on prior information, on data observable through date $t - 1$, *and* on $\beta(t - 1; \cdot, \cdot, \cdot)$, the vector $[\beta(t; i, \cdot, \cdot), u(t; i)]$ is taken to be a mixture of two jointly normally distributed random variables, both with mean $[\beta^*(t - 1; i, \cdot, \cdot), 0]'$ and with variance matrices $V(t; i)$ and $\pi_{11}^2 V(t; i)$, respectively; that is, the vector has p.d.f.[3]

2. In this model, the algebra of the "seemingly unrelated regressions" of econometric textbooks applies. Thus, even if the randomness is related across equations, if the same X's appear on the right-hand side of each equation and the prior has the same form in each equation, then analysis of the whole system reduces to equation-by-equation analysis. However, the prior that we consider is not symmetrical across equations.

3. Here and below, I will use the abbreviated notation $a(t)$ for $a(t; i, \cdot, \cdot)$ where there can be no ambiguity.

$$p[\beta(t), u(t; i)|t-1] = \pi_{10}\phi\left\{\begin{bmatrix}\beta^*(t-1)\\ 0\end{bmatrix}, V(t; i)\right\} + \left(1-\pi_{10}\right)\phi\left\{\begin{bmatrix}\beta^*(t-1)\\ 0\end{bmatrix}, \pi_{11}^2 V(t; i)\right\}, \tag{2}$$

where $\phi(a, b)$ is the p.d.f. of a normal vector with mean a and variance matrix b.

Conditional on data and prior information observable through $t-1$ *alone*, $\beta^*(t-1)$ is taken to be normally distributed with covariance matrix $W(t-1)$ and mean $B(t-1)$; that is, it is taken to have p.d.f.

$$q[\beta^*(t-1)] = \phi[B(t-1), W(t-1)]. \tag{3}$$

If π_{10} were zero or one, equations (1)–(3) would justify applying the Kalman filter to an observation on $X_i(t)$ to obtain a posterior distribution for β and u. With other values of π_{10}, the Kalman filter cannot be applied directly since the conditional distribution of $X_i(t)$ is nonnormal. However, the posterior distribution is still easily obtained by two applications of the Kalman filter. One applies it once conditional on the $V(t; i)$ covariance matrix, then again conditional on the $\pi_{11}^2 V(t; i)$ covariance matrix. The posterior p.d.f. on $[\beta(t), u(t; i)]$ is then a weighted sum of the two resulting normal posterior p.d.f.'s, with the weights given by the relative likelihoods of the observed $X_i(t)$ under the two normal prior distributions.

The posterior distribution on $\beta(t)$ generated by this procedure is, of course, itself a mixture of normals, not a normal distribution. If $\beta(t+1)$ were related to $\beta(t)$ by a linear equation with normal disturbances, the prior distribution on $\beta(t+1)$ would itself be nonnormal, and the Kalman filter would not be applicable at $t+1$. Actually, if the prior at $t=0$ is normal, the prior at $t=1$ would be a mixture of two normals, so that by conditioning on each normal component of the prior, Kalman filtering twice for each, we could obtain a new posterior that was a mixture of four normals, etc. However, with the number of normal components involved proliferating exponentially, this exact approach would be computationally intractable. A better approximate approach might be continually to keep track of the k most likely of the 2^t branches of the tree of normal components of the mixed posterior distributions, with k set at, say, four or sixteen. Or, instead, at each t one could convert the posterior for $\beta(t-k; i, \cdot, \cdot)$ conditional on data through $t-k$ to the normal distribution with corresponding mean and variance, treating the disturbances from $t-k+1$ to t exactly. What is actually done for this model is this latter approach with $k=1$, although a k of two or three would be feasible, at least as an experiment to check the sensitivity of results. One hesitates to work too hard at this since the mixture-of-normals assumption itself is an arbitrary convenience. A matrix t distribution would be more plausible, implying a continuous mixture of normals in place of a mixture of just two normals.

To summarize, the model assumes that $\beta^*(t-1)$ is a function of $\beta(t-1)$ such that it has a normal distribution with the same mean and variance as has $\beta(t-1)$, despite the nonnormality of the latter.

If we could represent this change in distribution by supposing that some sort of random noise were added to $\beta(t-1)$, it would be natural to think of this as simply nonnormal stochastic time variation in a. However, the nature of the change in distribution precludes its being characterized this way. The assumption is in fact unnatural, justifiable only as a convenient approximation. Note, however, that, because our uncertainty about $\beta(t-1)$ cumulates the effects of disturbances at many dates, our posterior for it is likely to be much closer to normality than is the conditional distribution for $\beta(t) - \beta^*(t-1)$. Treating the distribution of the former as approximately normal while carefully accounting for nonnormality in the latter is therefore justifiable as an approximation.

Note that we are in effect assuming that our posterior mean for $\beta(t-1)$ at $t-1$ is the same as our prior mean for $\beta(t)$. This makes the $E[\beta(t)|t]$ sequence a martingale. There would be no computational or conceptual difficulty with allowing a more general linear dependence of the prior mean for $\beta(t)$ on $\beta(t-1)$, and, indeed, in this and other models, Litterman and I have both experimented with specifications where

$$E[\beta(t) - \bar{\beta}|t-1] = \Theta E[\beta(t-1) - \bar{\beta}|t-1],$$

with Θ a scalar and $\bar{\beta}$ the prior mean vector. The best choice of Θ has always turned out to be close to one, however, so that, with sample sizes of the length actually available, there has seemed little advantage to freeing it to differ from one.

4.2.2 Initial Prior Mean

In the model discussed here, m, the lag length, is five—slightly over a year since I am using quarterly data. The vector $B(0; i, j, \cdot)$, the initial prior mean on $\beta(1; i, j, \cdot)$, is set to zero for $i \neq j$. The vector $B(0; i, i, \cdot)$ is given by

$$1.2679, -.3397, .0910, -.0244, .00654.$$

These numbers satisfy $B(0; i, i, s) = (1+\alpha)(-\alpha)^s$, which (if s is allowed to run to infinity instead of being truncated at five) defines the autoregressive coefficients for an ARIMA(0, 1, 1) process with moving average parameters $\alpha = 2 - \sqrt{3}$. It can be shown that this is the form of a unit-averaged Wiener process. Thus, the prior mean makes all elements of X behave like unit-averaged Wiener processes with no lagged cross-relations among components of X.[4]

4. Note that, in previous published work, prior means for BVAR models have generally made the components of X discrete-time random walks. The unit-averaged Wiener process prior (at least

4.2.3 The Initial Litterman Prior

The prior covariance matrix is built up by a sequence of modifications of an initial prior. The initial prior makes each scalar component of the $\beta^*(t; i, \cdot, \cdot)$ vector independent of all the others (i.e., it makes the covariance matrix $W[0]$ diagonal) and sets the variance according to

$$\sqrt{\text{Var}[\beta^*(0; i, j, s)]} = \frac{\sigma(i)}{\sigma(j)} \pi_1 \pi_2^{\delta(i,j)} \exp[-\pi_3 \log(s)], \quad j = 1, \ldots, k+1. \tag{4}$$

Here, $\sigma(i)$ is a parameter measuring the scale of fluctuations in variable i, taken in practice as the residual standard error from a univariate fifth-order VAR fit to the entire sample for $i \leq k$. For $j = k + 1$, there is only an $s = 1$ term, as the corresponding a is the "constant term" (here actually not a constant but time varying). For this term, $\sigma(j + 1) = 1/\pi_4$, another unknown parameter. The function $\delta(i, j)$ is the Kronecker delta, one for $i = j$, zero otherwise. Here, as elsewhere in this paper, the parameters π_i are "unknown constants." In principle, we should specify a prior over them to complete a Bayesian framework for inference. However, because doing so would be inconvenient, and because we expect that our prior on them would be uninformative (i.e., we do not know much about them a priori), we integrate over these parameters informally.

4.2.4 The Dummy Initial Observation

The range of differences in observed dynamic behavior for economic time series is fairly large, and, indeed, a reasonable prior specification for the standard error of $\beta^*(0; i, i, 1)$ is about 0.16. But then this component of uncertainty about β^* alone accounts for an implied standard error of forecast for $X_i(1)$ amounting to 16 percent of the initial level of X_i. Since the random components in the other elements of $\beta^*(0)$ are all independent of this one, they all serve only to increase the implied forecast errors. We are not in fact this uncertain about the accuracy of naive random walk forecasts (which is what our initial forecasts, based on prior means for β^*, will be). We are unsure of whether our prior means are exactly right, coefficient by coefficient, but we find it much more likely that the best forecasting model will be one that implies that naive no-change forecasts will be fairly good than that it will be one that implies that great improvements on a no-change forecast are possible. If coefficients deviate from their prior means, we expect that other coefficients

---

where the data have in fact all been collected as unit averages) is a notably more accurate naive standard, however. Observe that the Theil U's (for a definition, see the note to table 4.2 below) obtained by using the correct AR in place of a discrete random walk AR for a process that is actually a unit-averaged Wiener process would be, at forecast horizons 1–4, .933, .9732, .9832, and .9878.

will deviate in an offsetting way, with the result that naive no-change forecasts will still be fairly accurate.

To capture this aspect of prior beliefs, we need to introduce appropriate off-diagonal elements into $W(t; i)$ while leaving the diagonal elements relatively undisturbed. One easy way to do this is to introduce a "dummy observation" in which the prior is modified by feeding it into a Kalman filter that takes as observed data for $X(t - s)$, $s = 1, \ldots, m$, the actual m initial values of X from the sample and for $X_i(t)$, not the actual $X_i(m + 1)$, but instead the model's own forecast, based on the prior mean for β^*, of $X_i(m)$. The data in this dummy observation are weighted by a parameter π_5, which can be expected to be best taken to be near one if the variances of the $u(t; i)$ disturbances have been specified as near to the variance of forecasts from a naive random walk model. Because the Kalman filter finds that, with these artificial data, the prior mean generates perfect forecasts, the Kalman filter makes the posterior mean the prior mean. Only the variance matrix of the prior mean is changed. The change is of rank one and in practice turns out to have only modest effects on diagonal elements of W.

In most previous published work with BVARs, there has been a "sum-of-coefficients" modification to the prior. That modification can be characterized as a sequence of Kalman filtering operations indexed by $j = 1, \ldots, k$, in each of which $X(t - s)$ is set to zero for $s = 1, \ldots, m$, except for $X_j(t - s)$, $s = 1, \ldots, m$, all of which are set to one, while $X_i(t)$, the dependent variable, is set to one if and only if $j = i$. Because most economic time series are smooth, $X(t - s)$ and $X(t - s - 1)$ have similar values. Thus, the dummy initial observation used here is approximately a linear combination of the dummy observations used in imposing the sum-of-coefficients modifications. In practice, the dummy initial observation seems to reduce or eliminate the usefulness of sum-of-coefficients dummy observations. This point is substantively important because heavily weighted sum-of-coefficients dummy observations push the model toward a limiting form written entirely in terms of differences, which eliminates all long-run relations across variables. Putting the same point another way, the old sum-of-coefficients dummy observations pulled the model toward a form with as many unit roots as variables and no cointegration, while the current dummy initial observation pulls the model toward a form with unit-root nonstationarity in all variables without down-weighting the possibility of cointegration.

This dummy-initial-observation idea was discovered in the process of adapting BVAR methodology to a context where the number of series available for a model increases at several dates scattered through a historical sample. A natural approach to such a situation is to begin with a prior for a model with all the variables that will eventually be available, padding the data for variables that are initially unavailable with zeros. Applying the Kalman filter to the padded data is equivalent to applying it to a smaller model. The

prior means and variance matrix of the coefficients on unavailable variables are left unaltered by the Kalman filter when the data for them is set at zero. However, at the time when data on a series do become available, the prior shows an exaggerated version of the problem described above as motivating the dummy initial observation. The new data multiply large prior variances on individual coefficients to imply large forecast errors, and the uncertainty about coefficients on the newly entering variable shows no correlation with uncertainty about coefficients on the variables already in the model. We know in fact that the small model estimated up to this point is a good forecasting model, and the availability of data on a new variable has not made its forecast accuracy worse. To make the prior reflect this knowledge, a dummy observation, in which the prior mean coefficients at t are presented to the Kalman filter as making perfect forecasts for $t + 1$, is appropriate. The prior mean coefficients are all zero for the newly introduced variable, so the dummy observation expresses confidence in the small model estimated without the new variable. Covariances between coefficients are created by the dummy observation, so that deviations from zero in coefficients on the new variable imply likely offsetting changes in other coefficients to leave forecasts from the previously estimated small model fairly close to those of the expanded model.

4.2.5 Relative Tightness on Durable Goods Prices

There is an a priori basis for expecting that prices of durable goods frequently traded in open markets will follow stochastic processes well approximated as Wiener processes over short time spans. Thus, our prior mean is inherently more attractive for such variables than, for example, for GNP or unemployment. I therefore introduce into (4) an additional multiplicative factor

$$\pi_9 \mathrm{IDGP}(j), \tag{5}$$

where IDGP(j) is zero for variables that are not durable goods prices and one for variables that are. The latter are taken to be the value of the dollar, stock prices, and commodity prices. A case could be made for including three-month Treasury-bill rates and M1 in this list, but they were left out as not actually being prices of durable goods.

4.2.6 Inflation Neutrality

In theoretical models without money illusion, the price level can change without any effect on real variables. If the data show persistent changes in the price level, price-level neutrality implies certain restrictions on the coefficients of the model. In particular in a log-linear model, coefficients on the right-hand-side nominal variables should sum to one in equations with nominal dependent variables and to zero in equations with real dependent variables. The prior mean for the coefficients in this model already almost satisfies this restriction since only coefficients on own lags have nonzero means and these

sum almost exactly to one. To pull the prior in the direction of sticking with price-level neutrality, we can perturb it with a dummy observation in which current values and all lags of nominal variables are set at one while all other variables are set at zero. For the nine variables in this model, the nominal variables are the money stock, the price level, stock prices, and commodity prices.[5] The model uses such a dummy observation, scaled up by a factor of 2.0. The use of this dummy observation has little effect on the likelihood or the estimated coefficients, but, as with δ in the next section, there has been no systematic exploration of the parameter space allowing variation in this parameter.

4.2.7 Covariance Matrix of Disturbances

The upper-left diagonal component of the matrix $V(1; i)$ corresponding to all the β's is taken to be $\pi_6^2 W(0; i)$, that is, just a scaled version of the initial prior covariance matrix. However, the scale of this matrix is allowed to adapt over time to the observed squared errors in the model. The idea here is very close to that of the ARCH models pioneered by Engle (1982), but it differs in that, instead of variances being adapted to the sizes of past unobservable true disturbances (u and $\beta - \beta^*$ in our notation), they are adapted to the sizes of past actual forecast errors, that is, in our notation to sizes of

$$v(t; i) = X_i(t) - \sum_{j=1}^{k} \sum_{s=1}^{m} X_j(t - s)B(t - 1; i, j, s) + B(t; i, j + 1, 1).$$

The specification adopted here has the advantage that it makes the variance of disturbances at $t + 1$ known at t, allowing a single pass of the Kalman filter through the data to evaluate the sample likelihood function. More specifically, the scales of the $V(t; i)$ matrices are adapted to the recent history of forecast errors in all equations of the system according to the following scheme. Let $v^*(t; i; 0)$ be $v(t; i)$ divided by the model's implied variance for $v(t; i)$ conditional on the true disturbance matrix being $V(t; i)$, while $v^*(t; i; 1)$ is $v(t; i)$ divided by the model's implied variance for $v(t; i)$ conditional on the true disturbance matrix being $\pi_{11}^2 V(t; i)$. Then let

$$v^{**}(t; i)^2 = p_0 v^*(t; i; 0)^2 + p_1 v^*(t; i; 1)^2, \tag{6}$$

where p_0 is the posterior probability, given data at t, of the smaller variance normal component of the mixed distribution for the disturbance at t, and $p_1 = 1 - p_0$ is the posterior probability of the other component. If the model is correct, v^{**} should average out to about one.

Let

5. There could be some dispute over whether to treat the value of the dollar also as a nominal variable. It could behave either way, depending on how price-level changes are related across countries.

$$(7)\quad \tau(t;\, i) = 1 + \pi_7 \left[\pi_8 v^{**}(t;\, i)^2 + (1 - \pi_8) \sum_{i=1}^{k} \frac{v^{**}(t;\, i)^2}{k} - 1 \right] + \delta.$$

Then we take

$$(8)\quad V(t + 1;\, i) = \tau(t;\, i) V(t;\, i).$$

Thus, π_7 measures the overall responsiveness of forecast error variance to the magnitude of the current errors, and π_8 measures the relative weight on own errors versus system-wide average errors in making the adjustment. If $\delta = 0$ in this specification, each $v(t;\, i)^2$ is a martingale, but, since these terms are necessarily positive, they form martingales bounded below. Thus, with $\delta = 0$, the model implies that $v(t;\, i)$ converges almost surely to a constant. While this implication is perhaps no more unreasonable than the implications of martingale behavior for β itself (which we have imposed), experimentation with δ nonzero seems warranted. The current version of the model takes $\delta = 0.01$, which slightly improves fit over $\delta = 0$, but there has been no systematic exploration of the likelihood surface in δ as there has been for the π vector.

4.3 Model Fitting

What I have described above is an eleven-parameter[6] probability model for the nine quarterly observed time series in the model. A classically oriented statistician can ignore the Bayesian jargon in the model description, treat the β's as well as the u's as unobserved random disturbances, and interpret the π's as the model parameters. From this perspective, our estimation procedure is simply maximum likelihood (although, as mentioned above, since we add up individual equation log likelihoods to form the system likelihood used as the fit criterion, we are in effect assuming independence of all random disturbances across equations, a potentially unrealistic assumption).

My own view is that maximum likelihood is justifiable only as an approximation to a Bayesian procedure or as a device for summarizing a likelihood function. The most important single aspect of a likelihood function, at least if it has a well-defined peak, is its maximum. Nonetheless, we must bear in mind that the peak might not be well defined or that the shape of the likelihood may otherwise turn out to differ from the usual Gaussian shape. In practice, this means that, if likelihood turns out to be insensitive to some dimension of variation in β, we ought to verify that the implications of the model that are important to us—forecasts and policy analysis—are also insensitive to this dimension of variation. If not, results from several parameter settings should be studied.

6. This counts only the parameters π_1–π_{11}, not δ, the drift in the variance process, or the weight on the price-neutrality dummy observation. There has been no systematic exploration of the parameter space along these latter two dimensions, so eleven is probably the right measure for assessing how much overfitting is likely to have occurred.

The derivation and interpretation of the likelihood function for this type of model have been described in Doan, Litterman, and Sims (1984). The mechanics of likelihood maximization have been handled with a nonstandard hill-climbing routine, described in Sims (1986b). Because each function maximization is relatively expensive (involving a pass through the data with two Kalman filter applications at each sample point), it seemed important to use global information about the shape of the likelihood in deciding on each function evaluation. The program used, BAYESMTH, fits a surface to the observed likelihood values to generate a guess for the location of the function's peak. It is applied iteratively, with fifty to one hundred function evaluations used to obtain very rough convergence.[7] An advantage of the Bayesian hill-climbing routine is that it can be used at any iteration to generate a best guess at the shape of the likelihood, which is more important for inference than the precise location of the peak.

It is worth noting that, in 1986, when this form of the model was arrived at, the nine-variable version of the model could not be estimated on a PC. Programs were developed on a PC in a six-variable version of the model that took forty minutes to complete a single evaluation of the likelihood. Iterative maximization of the likelihood was carried out with likelihood evaluations on a Cray supercomputer, which could complete a likelihood function evaluation for the nine-variable model in about twenty seconds. Now, a 33MHz 486 PC can evaluate the likelihood for a full nine-variable version of the model in about ninety seconds.

4.4 Characteristics of the Fitted Model

In the rows labeled "87," table 4.1 shows the π vector that achieved the highest level of the likelihood function when the model was fit to data for 1949:III–1987:III in 1987:IV. (The data used are described in the appendix.) Observe that, with $\pi_{11} = 3.8$ and $\pi_{10} = 0.31$, the mixed distribution is notably nonnormal, with a fourth moment 2.43 times as large as that of a normal distribution with the same variance. This is about the same kurtosis as for a t-distribution with five degrees of freedom. Geweke (1992) finds similar kurtosis using a different form of nonnormality in modeling macroeconomic time-series data.

Parameter 7, at 0.25, is small enough to imply significant delay in the reaction of forecast error variances to the previous history of errors, but large enough to imply substantial adaptation within a year. Parameter 8, at 0.34, implies that more weight is given to system-wide average error than to an individual equation's own error in adapting forecast error variances to historical experience.

7. Iteration is ordinarily halted when, say, ten or twelve successive function evaluations produce cumulated change in the log likelihood of less than 0.5.

Table 4.1 **Likelihood-Maximizing π Vectors**

π Subscript	Model Version	π Value	Description of π
1	87	.17	Overall tightness
	6	.10	
	11	.24	
2	87	.19	Relative tightness on other variables
	6	.11	
	11	.17	
3	87	.90	Exponent for increase in tightness with lag
	6	.14	
	11	.92	
4	87	3.41	Standard error of constant term relative to σ (i)
	6	5.31	
	11	1.71	
5	87	.89	Weight on initial dummy observation
	6	3.03	
	11	2.21	
6	87	.09	Ratio of initial standard error of time variation
	6	.00	to initial prior standard error
	11	.10	
7	87	.25	Overall sensitivity of forecast error variance to
	6	.00	current error magnitudes
	11	.21	
8	87	.34	Relative weight on equation's own error size
	6	.00	vs. average of system error sizes in setting
	11	.32	variance evolution
9	87	.27	Relative tightness of the prior on durable
	6	1.28	goods price equations
	11	.12	
10	87	.31	Probability that the disturbance is drawn from
	6	.00	the normal component with larger variance
	11	.35	
11	87	3.76	Standard deviation of the more diffuse of the
	6	1.00	two components of the disturbance
	11	3.66	distribution, as a multiple of the standard deviation of the less diffuse

Note: A vector of 11 hyperparameters is displayed for each of three versions of the model. In each group of three numbers, the top one refers to a model fit in 1987:IV to data available then for 1949:III–1987:III, and the lower two refer to models fit in 1992:II to data available then for 1949:III–1992:I. The middle one was fit while parameters 7, 8, 10, and 11 were held fixed at the displayed values.

From parameters 1 and 2, we see that the prior standard deviation of the coefficient on the first own lag in each equation is about 0.2 and that coefficients on other variables are given prior standard deviations about 20 percent of the prior standard deviations on own lags. Parameter 3, close to one, implies that prior standard deviations on coefficients for lag s decline approximately as $1/s$. Parameter 6, at about 0.1, implies that the variance of a one-

period change in the coefficient vector is about 1 percent of its initial prior variance. The prior uncertainty about the coefficients is thus about the same as prior uncertainty about the parameter change over one hundred quarters.

The parameters that we have discussed to this point are all about the same in both the model fit in 1987 and the current update of that model. The remaining ones, parameters 9, 5, and 4, showed substantial changes with the update. As it stands, the model simply scales the prior covariance matrix by π_6 to obtain the covariance matrix of coefficient changes. Along these dimensions in which refitting has resulted in large changes, it is possible that the model should allow differences between the prior covariance matrix and the coefficient-change covariance matrix. Of course, it is also possible that these results simply reflect sample information. Note that, while the October 1987 stock market crash had occurred at the time of the 1987:IV model fitting, it was not in the data set on which the model was fit.

The biggest change is in parameter 9, which, in going from 0.27 to 0.12, implies a much tighter prejudice in favor of the random walk model for the durable goods price variables after refitting. This is in line with the fact, documented below, that the model's forecasts for these variables have shown little if any margin of superiority over those of a naive continuous-time random walk model. Parameters 5 and 6 show increased weight on the initial dummy observation and decreased prior variance for the constant term with refitting.

The differences in parameters between the 1987 version and the updated version of the model are enough to make modest but noticeable changes in model forecasts. For the 1982:II forecast, for example, the general shape of forecast paths for variables is little affected, but the level of long-run growth to which the forecast paths gravitate is affected. For variables not hit by parameter 9, these differences are on the order of 0.1–0.3 percentage points in the forecast annualized growth rates (and about the same magnitude in the forecast levels of interest rates). For the value of the dollar, stock prices, and commodity prices—the three variables hit by π_9—the forecast long-run annualized growth rates are affected by 1 or 2 percentage points.

The model with coefficient variation, nonnormality, and time-varying variances suppressed, reported in the middle rows of table 4.1, fits best with a tighter overall prior, relatively stronger prior restriction on cross-variable relations, weakly damped prior variances on longer lags, and small weight on the durable goods price restriction. The increased tightness in parameters 1 and 2, pulling all variables closer to the random walk model, roughly offsets the increased looseness in parameters 9, 3, and 4.

Imposition of these simplifying restrictions reduces likelihood for each equation, and the sum of the reductions in twice the log likelihood (which can be interpreted as measured in "chi-squared" units) is 1,146. This is a very large likelihood reduction, corresponding roughly to what would be produced by increasing forecast RMSE by 30 percent in every equation. Since the RMSE differences between the six-parameter and the eleven-parameter model

Table 4.2 **Theil *U* Statistics, 1949:III–1992:I**

		Quarters Ahead			
	Model Version	1	2	4	8
Treasury-bill rate	87	.9493	1.0295	.9786	.9845
	6	.9746	1.0362	1.0002	.9387
	11	.9682	1.0570	.9858	.9928
M1	87	.4546	.4354	.4252	.4265
	6	.4645	.4410	.4178	.3920
	11	.4489	.4305	.4185	.4203
GNP deflator	87	.3799	.3169	.2906	.2737
	6	.4350	.4053	.4184	.4347
	11	.3787	.3143	.2857	.2679
Real GNP	87		.6857	.6937	.6693
	6	.7388		.6557	.6047
	11	.7290	.6675		.6653
Business fixed investment	87	.8686	.8989	.9782	1.1132
	6	.8847	.9165	.9231	.8961
	11	.8533	.8655	.9272	1.0720
Unemployment	87	.7936	.8510	.9238	.9835
	6	.8110	.8827	.9612	.9615
	11	.7910	.8443	.9133	.9516
Trade-weighted value of dollar	87				1.2309
	6		1.0203	1.0585	1.1613
	11		1.0185	1.0664	1.1932
S&P 500 stock price index	87	.9126	.9197	.9217	.9470
	6	.9243	.9499	.9645	.9969
	11	.9234	.9444	.9593	.9932
Commodity price index	87	.8517	.9217	1.0308	1.2121
	6	.8128	.8951	1.0682	1.2043
	11	.8578	.9227	1.0021	1.0967

Note: In each group of three numbers, the figures shown correspond to the three parameter settings displayed in the corresponding groups of three numbers in table 4.1: an 11-hyperparameter model fit in 1987, a 6-hyperparameter model fit in 1992, and the 11-hyperparameter model fit in 1992. The Theil *U* is the ratio of root mean squared error (RMSE) of model forecasts to the RMSE of naive no-change forecasts for the same period.

shown in table 4.2 are not nearly this large, it is clear that much of the likelihood improvement comes from more accurate modeling of the evolution of forecast error variances in the eleven-parameter model.

4.5 Measures of Forecasting Performance

Table 4.2 shows how these differences in parameters affect model forecast performance. All the numbers in this table are Theil *U* statistics, meaning ratios of model RMSEs to naive no-change RMSEs. All the errors entering into these calculations come from using the model, with VAR coefficients recursively updated each quarter, to prepare forecasts through the sample. A

single consistent time series, constructed in 1992:II, is used for each variable, so that forecast errors measured here do not correspond exactly to the actual historical forecast errors. Data for periods after the forecast date affect the forecast for that date only insofar as they have influenced selection of parameters. The rows labeled 6 and 11 had parameters (π_j's, not VAR coefficients) fit to the full sample through 1992:I, while the rows labeled 87 had parameters fit to data through 1987:III only. Forecasts made and circulated regularly from this model from 1987:IV through 1992:II have all used the same set of parameter values, that corresponding to the "87" rows of table 4.2.

The 1987:IV–1992:I postsample period for the 1987 model is about 10 percent of the full sample. A 20 percent improvement in forecast accuracy over this period, with no deterioration in performance for the sample period, would improve the Theil U for the full period by about 2 percent. In column 1 of table 4.2, there are no improvements of this magnitude in the Theil U from the updated fit. As we move rightward along the columns, the forecasts whose accuracy is being measured overlap in time more and more, with the result that the sampling variation in the forecast accuracy measure, particularly over the short postsample period, increases substantially. The update of parameter estimates produces improvements in two-quarter forecasts of 2.7 percent in real GDP/GNP and 3.9 percent in investment. At the four-quarter horizon, there are improvements of 2 percent or more in these same two variables and also in commodity prices. At the eight-quarter horizon, GNP/GDP drops off the list, and the GNP/GDP deflator, unemployment, and the trade-weighted dollar are added. The only differences of 4 percent or more are for investment at the four-quarter horizon (5.5 percent), commodity prices at the eight-quarter horizon (10.5 percent), and stock prices at the eight-quarter horizon (4.9 percent) in favor of the 1987 version of the model). On the whole, it is clear that there is some gain from updating the fit, as would be expected, but that the model's out-of-sample forecast performance does not drastically contradict the 1987 parameter estimates.

The U statistics for the simplified six-parameter model show that, for most variables and time horizons, the eleven-parameter model performs better, but not by very much. The sharpest exception is for the GNP/GDP deflator. The smaller model is worse there by a large margin. At the eight-quarter horizon, the smaller model is 10 percent better for real GNP/GDP and 20 percent better for investment. Since this better performance for the smaller model is not matched at the shorter time horizons for these variables, it is hard to know what to make of it. It could be sampling variation, but it might also indicate a weakness in the larger model at long horizons for these variables.

Table 4.3 focuses on the performance of the 1987 version of the model for the postsample period, again using a single 1982:II data set rather than the historical sequence of regularly revised data series. The first three rows in each horizontal block of four rows compare the RMSE over 1988:I–1992:I for the 1987 model to the RMSE of alternative forecasting schemes: a naive no-

Table 4.3 **Performance in 1988:I–1992:I**

		Quarters Ahead			
		1	2	4	8
Treasury-bill rate	F/N	.8097	.9620	1.0444	1.1633
	F/S	.5890	.7660	1.0898	1.1497
	F/6	.7667	.8238	.8826	.8867
	F		.9686		
M1	F/N	.4702	.5165	.6272	.6527
	F/S	1.0676	1.1407	1.2289	1.1253
	F/6	.8387	.8283	.8414	.7093
	F	.0077	.0152	.0306	.0542
GNP deflator	F/N	.2246	.2140	.2063	.1157
	F/S	.4532	.5307	.5676	.3448
	F/6	.3983	.3342	.2793	.1298
	F	.0022	.0042	.0081	.0091
Real GNP	F/N	.7257	.7925	.7958	.9917
	F/S	.4456	.5407	.5065	.5734
	F/6	1.0236	1.0736	1.0230	1.0831
	F	.0043	.0087	.0147	.0278
Business fixed investment	F/N	.9077	.9019	.9840	1.5563
	F/S	.6251	.5606	.5374	.6133
	F/6	.8263	.7197	.6520	.8139
	F	.0153	.0245	.0426	.0844
Unemployment	F/N	.9419	1.0043	1.0177	.8768
	F/S	.5976	.5972	.6121	.5732
	F/6	1.1643	1.2738	1.4563	1.7785
	F	.0364	.6099	.1280	.1692
Trade-weighted value of dollar	F/N	1.0662	1.1994	1.4599	1.8827
	F/S	1.6418	1.5944	1.4049	1.0495
	F/6	1.0154	1.0362	1.1170	1.3192
	F	.0473	.0798	.1161	.1570
S&P 500 stock price index	F/N	1.0065	1.0050	.9396	.8444
	F/S	.9319	.9321	.9343	.8234
	F/6	.9958	.9000	.7980	.5859
	F	.0523	.0874	.1366	.1805
Commodity price index	F/N	1.0679	1.2901	1.5340	2.7498
	F/S	.4230	.3700	.3377	.4666
	F/6	.4600	.3388	.2599	.2875
	F	.0145	.0247	.0428	.0909

Note: In each horizontal block, F/N is the ratio of root mean squared error (RMSE) for the model to that for a naive no-change forecast, both for the forecast period 1988:I–1992:I. F/S is the ratio of RMSE for the model over 1988:I–1992:I to that for the model over 1949:III–1992:I. F/6 is the ratio of model RMSE to the RMSE for the 6-hyperparameter model displayed as the middle numbers in table 4.1. The row labeled F is the RMSE for the model over 1988:I–1992:I. The model is the version using hyperparameters chosen in 1987:IV to fit data for 1948:III–1987:III, except for the 6-parameter model used for comparison in the third rows. The 6-parameter model was fit to 1992:I data.

change model over 1988:I–1992:I, the 1987 model over 1949:III–1992:I, and the simplified six-parameter model (fit to data through 1992:I) 1988:I–1992:I. The fourth row shows the RMSE for the 1987 model over 1988:I–1992:I.

In certain senses, this period was unusually easy to forecast. Note that, for seven of the nine variables, the F/S rows are uniformly less than one. This means that, for these variables, RMSE was smaller in the postsample period than in the sample period. Yet, for most variables, comparison of the F/N row with the corresponding 87 row in table 4.2 shows little improvement, or even deterioration, in the model's performance relative to that of naive no-change forecasts. In other words, forecasting in this period was easy, by historical standards, for both the model and the no-change alternative, with the proportional improvement generally stronger for the naive models. This leaves the implications for evaluating the model ambiguous. The model has done "well" by historical standards, but such performance in a period when naive models are also doing "well" is weak support at best for the model.

There are two strong exceptions to this general conclusion. The GNP/GDP deflator forecasts for this period were better than their historical average, and they also improved relative to naive forecasts. This positive picture must be qualified by the fact that nothing within the model suggests that such a dramatically good performance is likely to be anything more than a random piece of good luck. The value of the dollar, on the other hand, was forecast with RMSE 60 percent worse than its historical average, in a period when naive forecasts at long horizons were getting better. The six-parameter model performed much better for this variable and this period.

For commodity prices, the naive model was much better than the model at forecasting horizons beyond one quarter. But this seems likely to reflect an unusual spate of good luck for the naive model since the 1987 model had better RMSE than its historical average and moving in the direction of the naive model by going to the six-parameter model produced drastically worse RMSE for this variable. A similar argument suggests that not too much should be made of the strong advantage of the naive model over the 1987 model at the eight-quarter horizon for investment.

The substantial advantage of the six-parameter model over the 1987 model at all horizons for unemployment is worrisome. The facts that the 87 model nonetheless showed substantial absolute improvement in RMSE over the sample period and that there is no correspondingly strong advantage of the six-parameter over the eleven-parameter model for the full sample in table 4.2 suggest that this, too, may well be a random fluctuation in the relative performance of the models.

4.6 Tracking the 1990–91 Recession

The model did not perform brilliantly in tracking the 1990–91 recession, but this may make it all the more useful to examine graphically how it be-

haved. This recession was mild, with no annualized growth rate for GNP/GDP much outside the ±3 percent range. Since the model's historical RMSE for GNP/GDP is about 0.5 percent in levels, and therefore 2 percent at an annual rate, it cannot be expected that the model would precisely track the path of GNP over such a mild fluctuation. On the other hand, the two quarters of negative GNP growth in 1990:IV and 1991:I were almost completely unanticipated by the model.

The pattern of changing forecasts over this period is largely explained by two factors: the forecast path for each variable adapts to the new initial conditions as errors are accounted for, and the steady drop in interest rates, largely unanticipated by the model, affects predictions for other variables over a two- to three-year horizon. Before looking at the data for actual forecasts over the span of the recession, I document the cross-variable effects of interest rate disturbances.

Figure 4.1 displays interest rate forecasts made in June 1992, using data on national accounts through 1992:I and some contemporaneous data on other variables. One forecast uses a 1992:II Treasury-bill rate of 3.687 percent, which is a guess, based on actual monthly data for April and May of the likely actual 1992:II value. The other two forecasts condition on 1992:II bill rates lower by 0.25 and 0.5 percent. The forecast is made by updating coefficients based on data through 1992:I for all variables, forecasting 1992:II using these

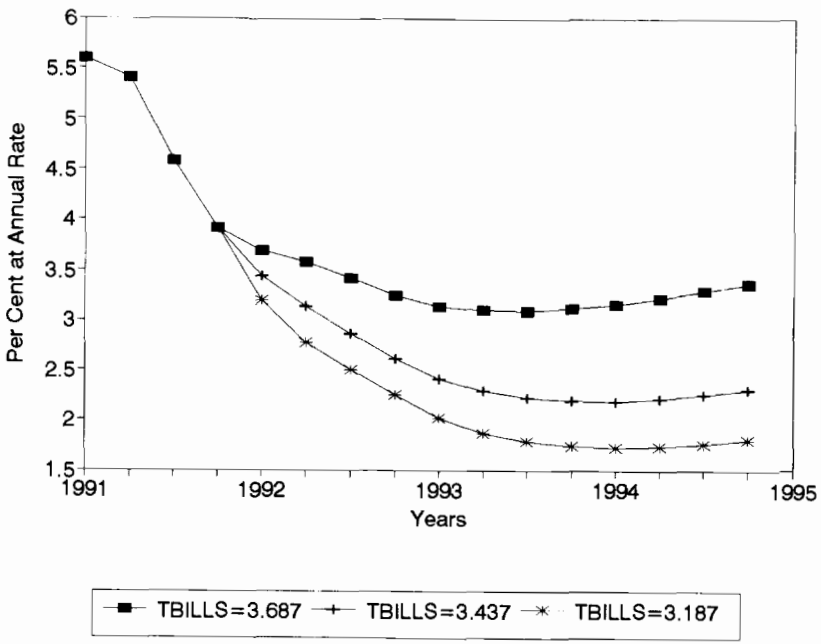

Fig. 4.1 Treasury-bill rate, June 1992 forecast

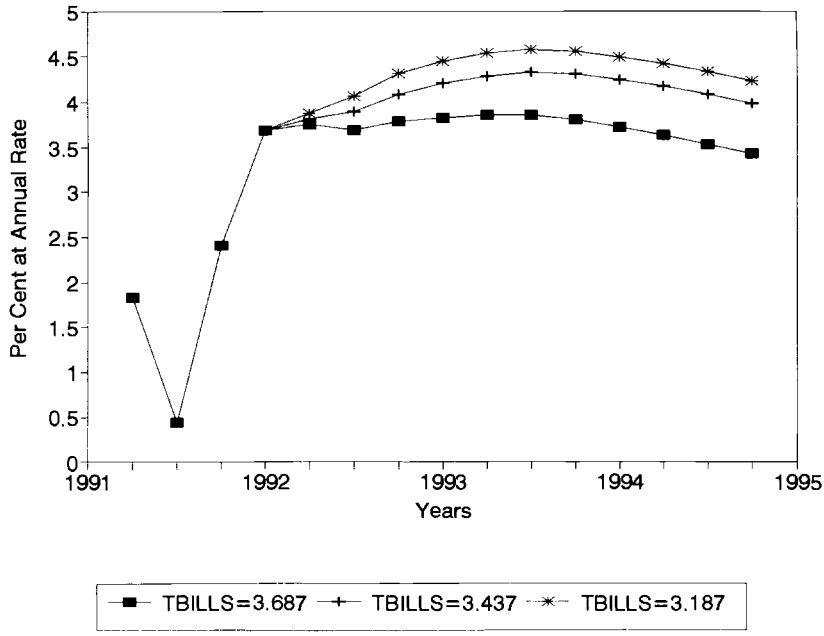

Fig. 4.2 Real GDP/GNP growth, June 1992 forecast

coefficients, replacing the forecast values for 1992:II for the bill rate, M1, the unemployment rate, the value of the dollar, and stock prices with "actual" values based on monthly data for the first two months of the model, updating the coefficients again, treating this mixed vector of forecast and actual values for 1992:II as if it were a new data point, then projecting from 1992:III onward with this final set of updated coefficients. The parameters of the model are those obtained from 1992 updates of the likelihood (the "11" rows in tables 4.1 and 4.2), not those of the 1987 model. Note that these interest rate forecasts diverge and that the paths are spread apart by much more than their original 0.5 percent dispersion after a year or two. Note also that nonlinearity shows itself clearly—since the Treasury-bill rate enters the model untransformed, an initial 0.25 percent perturbation in its path would in a linear model produce exactly half the perturbation in the remaining forecast path that an initial 0.5 percent would. Here instead the 0.5 percent initial perturbation produces considerably less than double the effect of a 0.25 percent perturbation after the initial period.

Figure 4.2 shows that the lower interest rates imply a forecast of more rapid output growth. Fig. 4.3 shows that the lower interest rates imply only slightly different GDP deflator inflation forecasts. The forecast inflation is lower with lower interest rates, but by a small enough amount that most of the drop in nominal rates still translates into a lower real rate. Fig. 4.4 shows that the sign

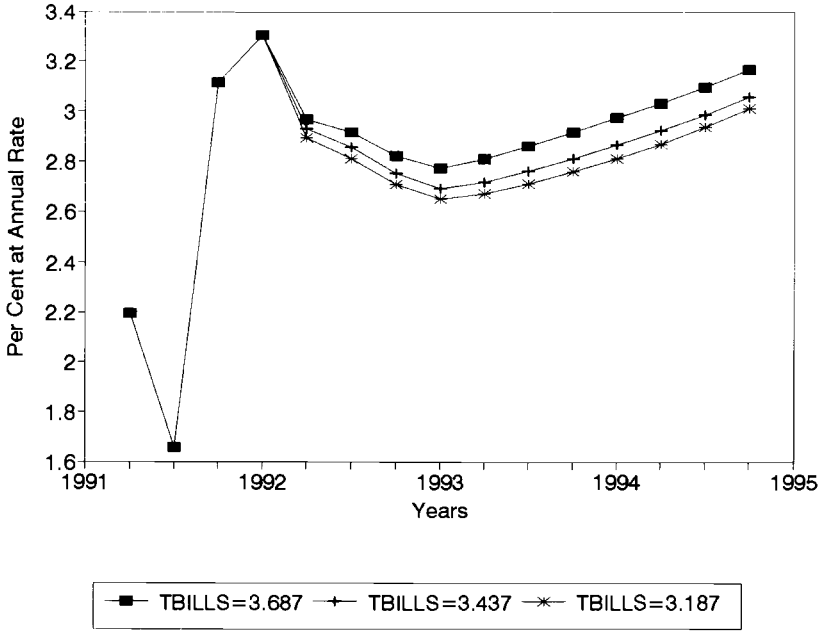

Fig. 4.3 GNP/GDP deflator inflation, June 1992 forecast

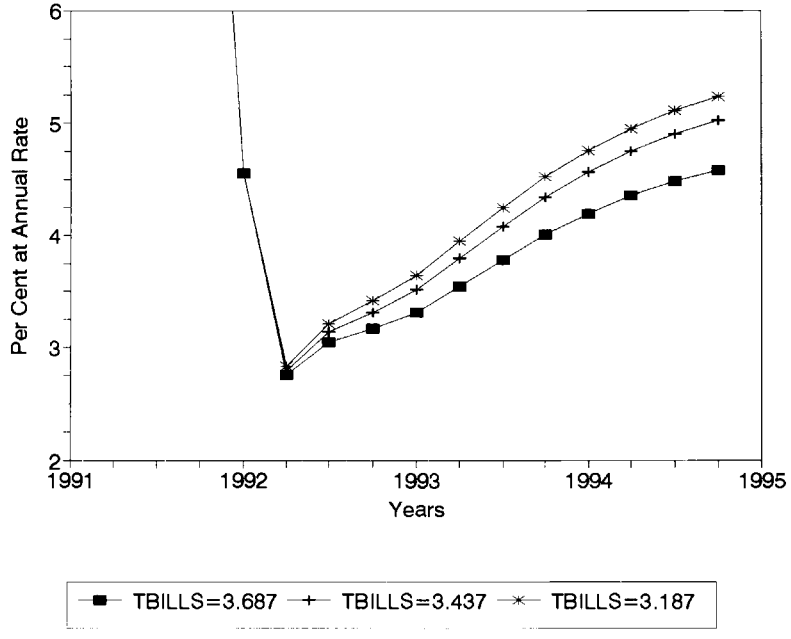

Fig. 4.4 Commodity price inflation, June 1992 forecast

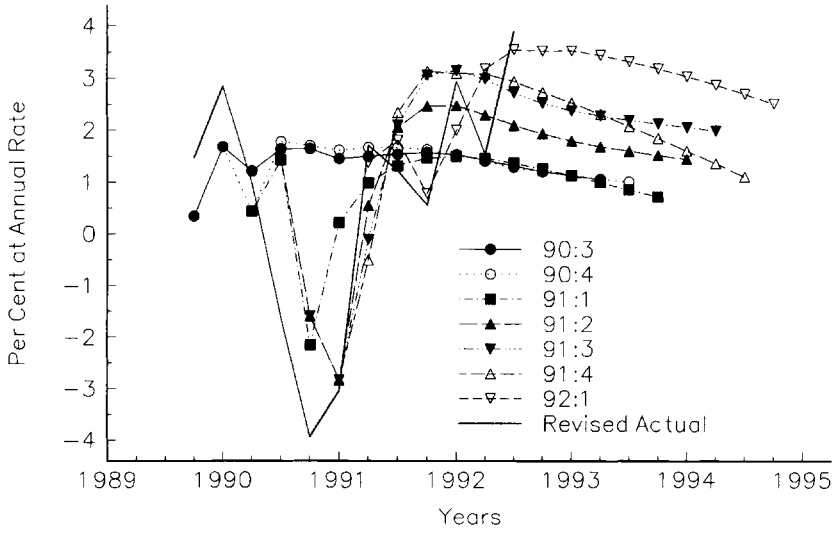

Fig. 4.5 Real GNP/GDP growth forecasts

of the effect on inflation in commodity prices is the opposite of that on inflation in the GDP deflator and that the magnitude is larger. (Note, however, that the commodity price index is much more volatile than the GDP deflator.)

This pattern of results is similar to the pattern that I noted in an earlier paper (Sims, in press) analyzing data across several countries with simpler VAR models. It raises interesting questions of interpretation, discussed in the earlier paper, that I leave aside here.

Figure 4.5 shows successive forecasts of GNP/GDP growth rates over the recession period, together with the actual growth rates for the period as shown in the revised GDP data available in December 1992. Unlike the forecast errors reported in the tables, these are actual historical forecasts, so the effects of data revisions can be seen in the plots. Each plotted line shows actual data for three quarters before the first quarter for which there were no data on GNP at the time of forecast, together with forecast values for the twelve subsequent quarters. In 1990:IV, the forecast (which used data through 1990:III for national accounts but some current data on interest rates, unemployment, money, the exchange rate, and stock prices) still showed no negative growth, although the projected positive growth rate of about 1.5 percent was low by historical standards. The 1990:IV GNP data were enough to pull the forecast growth for 1991:I down to less than 1 percent at an annual rate, but this left an error, compared to the actual negative growth of nearly 3 percent, nearly as large as for 1990:IV. The 1991:II and later forecasts, however, have been fairly well on track for the pattern of slow recovery since then.

Figure 4.6 shows how these patterns appear on a graph in levels. The fact

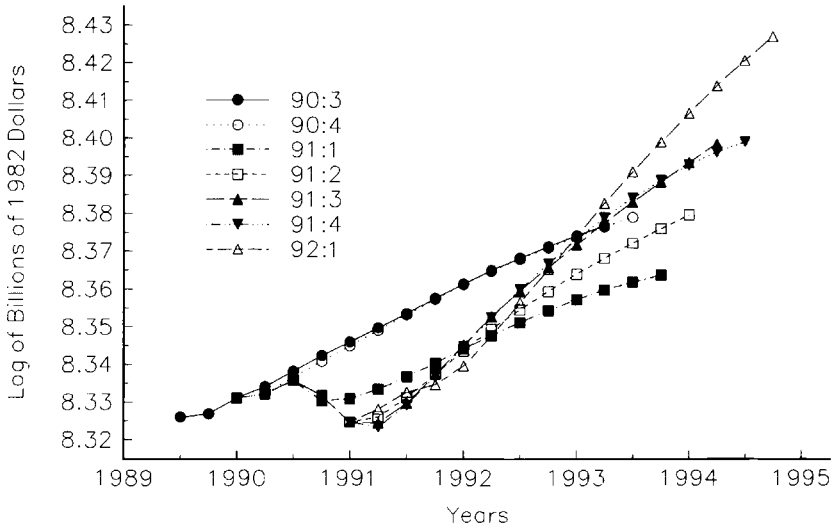

Fig. 4.6 Real GNP/GDP level forecasts

that the prerecession forecasts were for slow growth did not prevent them from being substantially mistaken about the 1991–92 levels of GNP. If growth proceeds as the model projects through 1992, GDP will be back up to its 1990:IV projected path around the end of 1992.

Figure 4.7 shows the sequence of interest rate forecasts. In every quarter except 1991:III, the interest rate fell by substantially more than the model had anticipated. On the basis of figures 4.1–4.4, we should expect that these interest rate shocks should increase forecast growth rates for output, slightly decrease forecast GDP deflator inflation, and increase forecast commodity price inflation over a two- to three-year horizon. We can see the corresponding increasingly optimistic long-run output growth forecasts in figures 4.5 and 4.6. Figures 4.8 and 4.9 show that, indeed, long-run GDP deflator inflation forecasts shifted downward and long-run commodity price inflation forecasts shifted upward. Note also the substantial effect of the data revision and switch to GDP rather than GNP accounting between the 1991:IV and the 1992:I forecasts. From figure 4.8, the "actual" inflation rate for 1991:III used in forming the 1991:IV forecast can be seen to be more than a percentage point above that used in forming the 1992:I forecast.

4.7 Conclusion

This model represents a further step in a research program attempting to bring into the realm of explicit probabilistic theory more of our uncertainty about the way the economy works. The model has been used for forecasts

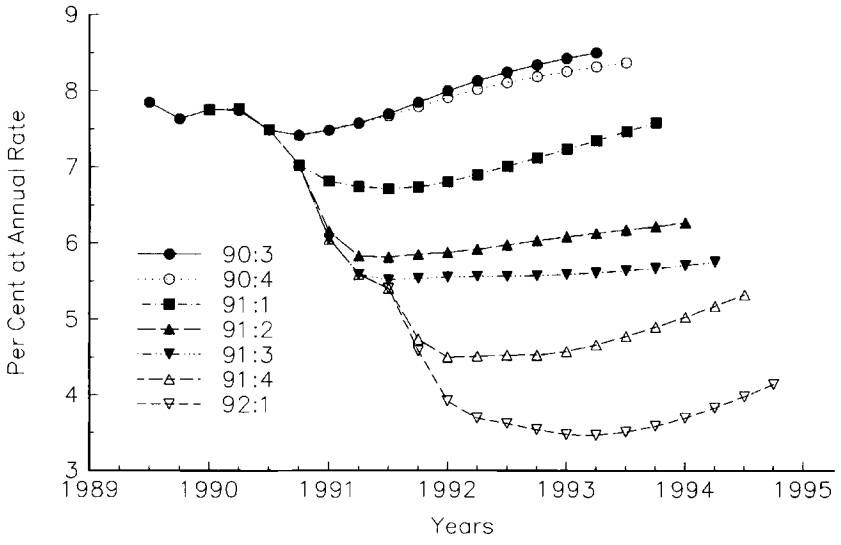

Fig. 4.7 Treasury-bill rate forecasts

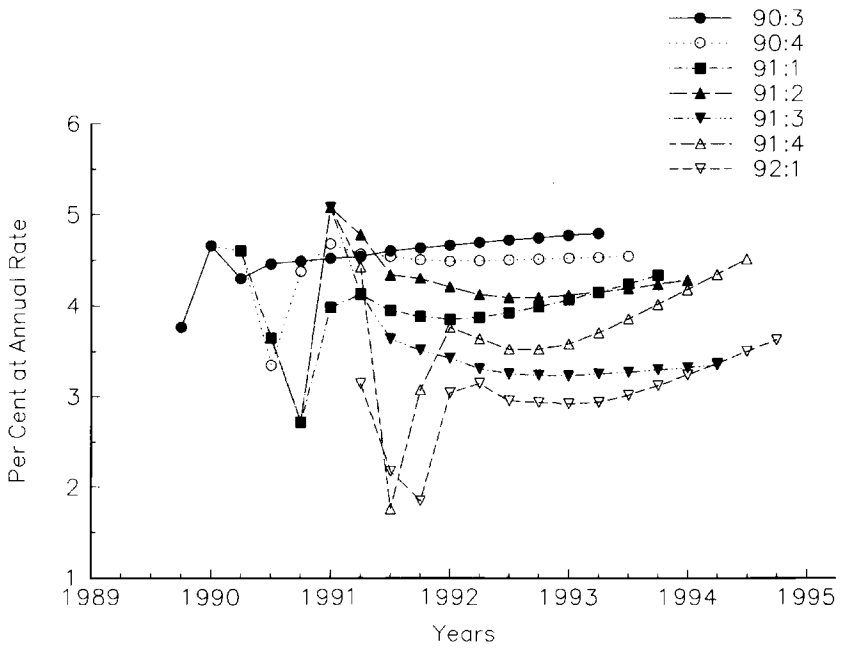

Fig. 4.8 GNP/GDP deflator inflation forecasts

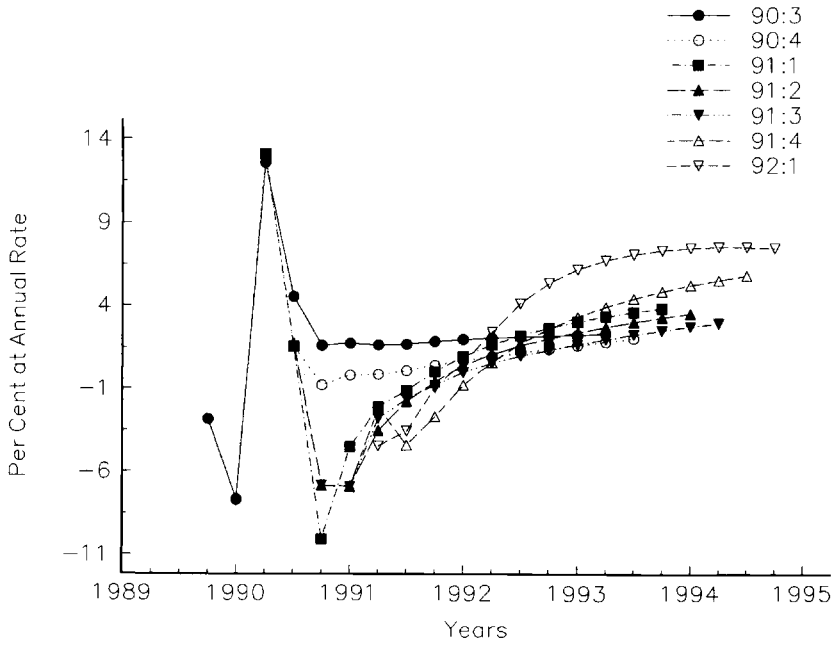

Fig. 4.9 Commodity price inflation 1992 forecasts

with the same parameters (the π vector) since mid-1987 and has performed reasonably well. It is a part of a sequence of models that have been used for forecasting since 1980, all of which have made forecasts without any add factors or ad hoc adjustments in response to current data over the entire period of record.

The form of the model has some implications for developments in macroeconomic theory that aim at explaining observed data. The model has substantial time variation in its coefficients, which is essential to generating good forecasts for some variables. Theories that imply linear models with fixed coefficients will therefore inevitably fall short. Rational expectations theorists, who have taken the lead in developing explicitly stochastic models, have not yet generated econometrically usable structural models capable of fitting a world of stochastically drifting parameters.

Recently, a number of authors (e.g., Bernanke 1986; Blanchard and Watson 1986; and Sims 1986a) have explored the use of convenient schemes for interpreting stationary VAR models. It is either discouraging or challenging, depending on your point of view, to note that, just as tools for convenient identification of stationary VAR models begin to be widely used, evidence emerges that stationary VAR models are inadequate. The problem of generat-

ing convenient identification schemes for the nonlinear, nonnormal model laid out in this paper appears quite difficult.

Appendix
Data Description

Treasury-bill rate. Three-month Treasury-bill rate, auction average.

M1. Pre-1959 data on M1, spliced together with more recent official data. For this series as for others described below as spliced, the splicing is done simply by scaling the earlier data to match the level of the more recent data at the date of switch.

PGDP. GDP deflator, 1987 = 100, seasonally adjusted. Spliced to earlier data on the GNP deflator at 1959:I.

GDP87. GDP, 1987 prices, seasonally adjusted. Spliced to earlier data on real GNP at 1959:I.

BFI87. Business fixed investment in 1987 prices, seasonally adjusted, from the GDP accounts. Spliced to earlier data from the GNP accounts at 1959:I.

UNEMP. Unemployment rate, civilians aged twenty and over, seasonally adjusted.

DOLLAR. Federal Reserve Board trade-weighted index of the value of the U.S. dollar, 1973 = 100.

STOCKS. The Standard and Poors 500 stock index, 1941–43 = 10.

PCOMM. Sensitive intermediate and crude producer prices index, January 1948 = 100 (U.S. Department of Commerce, *Business Conditions Digest,* ser. A0M098).

References

Bernanke, B. 1986. *Alternative explanations of the money-income correlation.* Carnegie-Rochester Policy Series on Public Policy. Amsterdam: North-Holland.

Blanchard, Olivier, and Mark Watson. 1986. Are all business cycles alike? In *The American business cycle,* ed. R. J. Gordon. Chicago: University of Chicago Press.

Doan, T., R. Litterman, and C. Sims. 1984. Forecasting and policy analysis using realistic prior distributions. *Econometric Reviews* 3 (1): 1–100.

Engle, Robert F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50: 987–1008.

Geweke, John. 1992. Priors for macroeconomic time series and their application. Institute for Empirical Macroeconomics Discussion Paper no. 64. Minneapolis: Federal Reserve Bank of Minneapolis.

Litterman, Robert B. 1986. Forecasting with Bayesian vector autoregressions—five years of experience. *Journal of Business and Economic Statistics* 4 (1): 25–38.
McNees, Stephen K. 1986. Forecasting accuracy of alternative techniques: A comparison of U.S. macroeconomic forecasts. *Journal of Business and Economic Statistics* 4 (1): 5–24.
Sims, Christopher. 1986a. Are forecasting models usable for policy analysis. *Quarterly Review of the Federal Reserve Bank of Minneapolis* 10 (Winter): 2–16.
———. 1986b. BAYESMTH: A program for multivariate Bayesian interpolation. Center for Economic Research Discussion Paper no. 234. University of Minnesota.
———. In press. Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review.*

Comment

Pierre Perron

It is a humbling experience to be asked to comment on such a paper. Forecasting economic time is surely not an easy task, and relatively few succeed at providing a useful product that is based on scientific principles and is free of the so-called add factors. Christopher Sims has been the intellectual leader of a class of forecasting models that, indeed, can claim such success. This paper provides an overview of the main features of the model as it now stands as well as several assessments of the quality of recent forecasts. Being neither an expert on the topic of forecasting nor particularly well trained in the Bayesian tradition, I will restrain myself to some general remarks that I hope will help the reader better understand and appreciate some issues underlying this methodology. These comments pertain to the following topics: the Bayesian interpretation, the treatment of trends, and the use of the mixture of normal distribution and its implications.

The Bayesian Interpretation

When reading a discussion of a paper that uses Bayesian tools, the reader often expects the discussant to probe or question the priors. I will refrain from doing so. In a sense, I am inclined to think that little can be gained from discussing whether one prior is better than another. What appears more important is to question the robustness of the results to reasonable changes in the prior specifications. This is, in principle, desirable, but in a project of this magnitude it appears difficult to ask the investigator for a full sensitivity analysis. In effect, I have nothing against the priors; none seem particularly unreasonable or, again, particularly undebatable. Many are imposed explicitly to make the model tractable. This is, however, fair game and does not appear any less a flaw than other simplifications found in alternative methodologies.

Pierre Perron is associate professor of economics at the University of Montreal and a research associate at the Centre de Recherche et Développement en Economique.

What strikes me, however, is that the model and the way in which it is estimated can hardly be said to be more Bayesian than any other popular approaches such as the standard VAR, the structural VAR (e.g., Blanchard and Watson 1986), the error correction type model of Stock and Watson (1991), or even, if I dare say so, the traditional simultaneous equations models of the Cowles Commission type (e.g., Fair 1984).

The basic starting point of the model is a data-generating process (DGP) of the general form

$$x(t) = f[x(t-1), \ldots, x(t-m); \theta(t)], \tag{1}$$

where $x(t)$ is the $k \times 1$ vector of variables in the system (the nine macroeconomic time series). In application, f is specified as a VAR of order m with time-varying parameters:

$$x_i = \sum_{j=1}^{k} \sum_{s=1}^{m} x_j(t-s)\beta(t; i, j, s) + \beta(t; i, j+1, 1) + u(t; i), \quad i = 1, \ldots, k. \tag{2}$$

So $\theta(t)$ contains here all the β's. The parameters $\theta(t)$ are specified to evolve according to a general function of the form

$$g[\theta(t)|\Phi(t-1)] = g\{\theta(t); \theta(t-1); [x(t-s), s \geq 1]; \mu\}, \tag{3}$$

where $\Phi(t-1)$ is the information set dated time $t-1$. Here the vector μ are some "hyperparameters" that govern the evolution of the parameters of the model. The vector of parameters $\theta(0)$ is also treated as a random variable with a marginal probability density function of the form $h[\theta(0), \gamma]$. Here μ and γ are the π's in the notation of the paper (eleven of them in all). Hence, the likelihood of the data is

$$p[x(t); t = 1, \ldots, T|x(-m-1), \ldots, x(0); \mu, \gamma]. \tag{4}$$

The model is estimated by maximizing the likelihood function with respect to μ and γ (more precisely, the method of estimation is quasi maximum likelihood since all k equations are estimated separately; as stated, it would be maximum likelihood estimation with no correlation across errors in the different equations). I see very little that is explicitly Bayesian in this approach; no priors are imposed over the π's to obtain the posterior distribution.

Sims (1991) argues that what makes this modeling approach explicitly Bayesian is the presence of the probability density function $h[\theta(0), \gamma]$ for the vector of initial conditions for the time paths of the parameters $\theta(t)$. The argument is that maximum likelihood estimation of γ will not yield a consistent estimate, thereby invalidating the standard justification for the classical approach (essentially because the accumulation of information as additional data become available does not add information about γ given that the initial conditions have transient effects). I think that such an argument is open to debate.

This point can be illustrated as follows. In a state space model with stationary variables, one can use the ergodic distribution of the parameters to initialize the system. When the time paths of the variables are nonstationary, this is no longer feasible, and one must either condition on some initial values or specify probability distributions for them. Sims's (1991) arguments imply that doing the latter necessarily makes the model Bayesian. I do not think that a classical interpretation of this modeling is strained. The initial distribution need not be viewed as imposing some priors; rather, it can be viewed simply as a convenient way to allow for randomness in the initial condition. The fact the maximum likelihood estimation on these parameters is inconsistent is also not much of a problem in a classical approach given that these initial conditions are often viewed as nuisance parameters, that is, parameters that are not the object of direct inference. In any event, it appears less strained to give a classical interpretation to the imposition of a probability distribution on the initial conditions than it is to give a Bayesian interpretation of the estimation method obtained without specifying prior distributions on μ.

The many priors imposed and discussed at length to justify them are ways to impose restrictions on the unrestricted time-varying parameters VAR given by (2). After all, this VAR is much too general, and some restrictions need to be imposed. This is no different than for any other forecasting model.

The final product is basically a constrained VAR where the constraints are highly nonlinear. Hence, the priors can be viewed as simply imposing a priori restrictions as in any other more "structural" model. The basic difference here from, say, the Cowles Commission approach is that the restrictions are "justified" using prior distributions that are themselves justified by extraneous empirical results (e.g., the plausibility of unit roots characterizing the time series) instead of being based primarily on theoretical arguments.

Let me stress that none of the comments above are meant to carry any negative connotations. On the contrary, I view it as quite an achievement when one is able to end up with a successful model with only some eleven free parameters by imposing restrictions that do not appear unbelievable. My view is basically that the priors act as alternative justifications for parameter reduction efforts that are present in all methods. The advantage here may lie in the highly nonlinear aspects of the restrictions that are eventually achieved.

The Treatment of Trends

One of the priors imposed is that the parameters are such that the data series have a prior mean of independent integrated processes. Some have argued that this is basically equivalent to first-differencing the data prior to the analysis. This implication is false, as argued in Sims (1991), because only the mean of the prior is centered on the unit root but variability is allowed. Note, however, that such centering on the unit root creates potential problems associated with the probability mass put on explosive processes. It might be more appropriate

to truncate or downweight the possibility of explosive roots relative to stationary ones.

The main point that I wish to convey, however, is that there is a more subtle way in which the structure of the model implies *fitted* unit root processes for any series that seems to be characterized by a nonstationary mean (e.g., real GNP, M1, the GNP deflator, stock prices, etc.). This is linked to the treatment of trends or, indeed, their absence. The basic fact to recognize is that no time trends are included as regressors in the estimated VAR as described (prior to the imposition of restrictions) by (2). My argument is that leaving out time trends (of any variety, linear or segmented, with constant or random coefficients) implies fitted series with unit roots. As much as Sims does not appear particularly to be a proponent of unit roots—see, for example, Sims (1988) (and I am not either; see Perron [1989])—the absence of trends in the estimated model precludes having series estimated to be characterized by "trend-stationary" processes (which we may view as equivalent to using differenced data).

To see how the argument goes, it is useful to consider a standard VAR with time-invariant parameters (for a more detailed treatment, see also Campbell and Perron [1991]). Consider a DGP of the form

$$x_t = \mu + \delta t + Z_t, \quad A(L)Z_t = e_t \sim N(0, \Sigma), \tag{5}$$

with $A(L)$ a matrix polynomial of order m in the lag operator L. We can write (5) as

$$A(L)x_t = A(1)\mu + \Phi\delta + A(1)\delta t + e_t, \tag{6}$$

where $\Phi = \sum_{i=1}^{m} iA_i$ is the mean lag matrix. Alternatively, denote $A(1) = \Pi$ following Johansen's (1988) notation for cointegrated systems:

$$\Delta x_t = \mu^* + \Pi x_{t-1} - \Pi\delta(t-1) + \text{lags}(\Delta x_t) + e_t \tag{7}$$

or

$$\Delta x_t = \mu^* + \Pi Z_{t-1} + \text{lags}(\Delta x_t) + e_t. \tag{8}$$

Note that, in general, the vector Z_t may contain elements with unit roots as well as stationary processes. It depends on the values of the matrices in the lag polynomial $A(L)$. To see that the fitted series will behave as unit root processes if the estimated model does not contain trend regressors, consider the following. A VAR with no time trends will be misspecified unless $\Pi\delta = 0$. In general, with a cointegrated system, $\Pi = \alpha\beta'$, where α and β are $k \times r$ matrices, r being the number of cointegrating vectors present and β being the matrix whose columns are the cointegrating vectors. Hence, $\Pi\delta = 0$ if and only if $\beta'\delta = 0$; that is, the cointegrating vectors that annihilate the nonstationarity in the noise component ($\beta'Z_t \sim I[0]$) also annihilate the nonstationarity in the trend component. Such a condition is often referred to as "deterministic cointegration."

The condition $\beta'\delta = 0$ may be one that is natural to impose, but it precludes the possibility of variables with a stationary noise component. Suppose that one of the variables (say the jth) is trend stationary, that is, $\delta_j \neq 0$ and $Z_{j,t} \sim I(0)$; then one of the rows of β is $e_j = (0, \ldots, 1, 0, \ldots, 0)$ (with a one in the jth position). Here, the unit vector is a (trivial) cointegrating vector. However, one of the conditions for the nonmisspecification of a VAR with no trends as regressors is that $e_j'\delta = \delta_j = 0$, that is, that the series is trendless. Accordingly, the model is misspecified if there is a series with a nonzero trend and a stationary noise component. The effect of this possible misspecification is to bias the parameters in such a way that the fitted values imply the presence of a unit root, and this bias does not vanish in large samples (for a proof in the univariate case, see Perron [1988]).

The example given above shows the importance of the treatment of trends. While I considered only simple linear time trends in the context of VAR with constant coefficients, the same principle applies in a more general context. The presence of trend regressors is necessary if trending series with stationary noise components are to be entertained as a possibility. I do not suggest the inclusion of linear time trends with constant coefficients in each equation. An interesting extension would involve the inclusion of time trends in the general VAR (2) with time-varying coefficients as well. The changes in these coefficients may be infrequent, and the prior distribution of mixed normality for the coefficients is particularly well suited for this purpose (I discuss this further below).

This increased generality would not come without drawbacks, however. It has been documented that the presence of time trends in estimated autoregressions creates, in general, a large downward bias on the sum of the autoregressive coefficients (see, e.g., Andrews 1991). Note, however, that this bias, unlike the ones created without trends, vanishes in large samples. In any event, asymmetrical priors could correct this bias. Such an extension may also introduce complexities in the specification of the priors. For example, flat priors on the trend coefficients and the autoregressive parameters are no longer adequate. Consider, for example, a simple univariate AR(1) such that $x_t = \mu + \delta t + Z_t$ with $Z_t = \alpha Z_{t-1} + e_t$. Such a process can be written as

$$x_t = (1 - \alpha)\mu + \delta\alpha + (1 - \alpha)\delta t + \alpha x_{t-1} + e_t \tag{9}$$

or

$$x_t = \mu^* + \beta t + \alpha x_{t-1} + e_t, \tag{10}$$

where $\mu^* = (1 - \alpha)\mu + \delta\alpha$, and $\beta = (1 - \alpha)\delta$. Note that, when $\alpha = 1$, $\beta = 0$, and a good prior should reflect this relation. Accordingly, priors should not be imposed on α, β, and μ^* but directly on α, δ, and μ. This may be more difficult to implement in practice. For similar arguments cast explicitly in a Bayesian framework, see Uhlig (1991).

The Mixture of Normals Prior

In his concluding comments, Sims argues that "the model has substantial time variation in its coefficients, which is essential to generating good forecasts for some variables." The importance of time variation in the coefficients is indeed highlighted throughout the paper. I certainly agree with this interpretation and the fact that the introduction of time variation in the parameters is a major step forward. However, in this model, time variation is associated with another modeling device, namely, specifying the form of the distribution of disturbances as a mixture of two normals. I think that such an assumption (or prior) is not simply "an arbitrary convenience" but rather a potentially equally important feature of the model.

The mixture of normal assumption for the distribution of the disturbances has the following form for the vector $[\beta(t), u(t; i)]$:

$$(11) \quad p[\beta(t), u(t; i)|t-1] = \pi_{10}\phi\left\{\begin{bmatrix}\beta^*(t-1)\\ 0\end{bmatrix}, V(t; i)\right\} + (1-\pi_{10})\phi\left\{\begin{bmatrix}\beta^*(t-1)\\ 0\end{bmatrix}, \pi_{11}^2 V(t; i)\right\},$$

where $\phi(a, b)$ is the normal density function with mean a and variance b. This specification states that $[\beta(t), u(t: i)]$ has mean $[\beta^*(t-1), 0]$, where $\beta^*(t-1)$ is also specified to follow a specific distribution. Its variance, however, is $V(t; i)$ with probability π_{10} and $\pi_{11}^2 V(t; i)$ with probability $(1-\pi_{10})$.

While the use of mixtures of normal distribution, and other non-Gaussian distributions in general, for disturbances is rather new in econometric modeling, it has some history in the statistics literature (see, in particular, Kitagawa [1987] and the comments related to that paper, esp. Martin and Raftery [1987]). The main motivation for its use is the modeling of structural changes (and outliers) in a time series of data. To see why this is so, consider the case where, in (11), π_{10} is small and π_{11} is large (substantially larger than one). In this case, the disturbances are drawn from the low-variance normal distribution most of the time, but, occasionally, a disturbance is drawn from the normal distribution with high variance. This effectively introduces a fat-tailed type behavior for the disturbances. To see why this can be useful in the analysis of structural change, note first that the β's follow martingale-like paths and that the x's are, in general, integrated as well (in the sense that the errors $u[t: i]$ have a permanent effect on the level of the series $x[t]$). In this case, a disturbance $u(t; i)$ drawn from the high-variance component will create a pattern similar to a structural change in the level of the x's (since the event is relatively rare and of a different order of magnitude than the disturbances that are drawn most of the time). Similarly, a disturbance to the coefficients $\beta(t; i, j+1, 1)$ (in the notation of [2]), that is, in the constants of each equation, will create a pattern similar to a structural change in the rate of growth of the

series. Disturbances from the high-variance component to other coefficients create major changes in the autoregressive coefficients that are more difficult to interpret but that could include structural changes in cointegrating vectors. Note that, if trends are included in the specifications to allow "trend-stationary" series, similar disturbances to the u's and the constants become, respectively, outliers and changes in level while structural changes in slopes would be associated with draws from the high-variance component for the coefficients on the time trends.

Many recent papers have argued that structural changes of the type described above are likely to be important ingredients in the characterization of many economic time series (e.g., Perron 1989, 1990; Hamilton 1989; Chen and Tiao 1990; Gregory and Hansen 1992; and others). A recent study by Park (1992) also extends Kitagawa's (1987) framework explicitly allowing mixtures of normals and shows its relevance in characterizing some aggregate economic time series involving a one-time change in slope. For these reasons, I think that the mixture of normals specification is likely to be a key ingredient in the success of this methodology. It would indeed be of substantial interest to report in some future work the time path of the coefficients and the implied decomposition into trend and noise components.

Several comments stand out from the structural change interpretation discussed above. First, let me reiterate the point made in the last section about the potential importance of including trend regressors to allow for the presence of series with a noise component that is stationary (in the sense of having no unit root but not excluding possible changes in the autoregressive coefficients). When allowance is made for the possibility of structural changes in slope and/or intercept, such a generalization becomes even more relevant (see Perron 1989).

A comment specifically directed to the specification of the mixture of normals (11) is the following. Note that there is only one parameter, π_{10}, measuring the probabilities associated with each component of the mixture; accordingly, the probability of drawing a disturbance from the high-variance component is the same for all elements of the vector $\beta(t)$ (i.e., the drift and the autoregressive coefficients) and for the errors $u(t; i)$. Under the interpretation discussed above, this implies the same probability of occurrences for changes in level, in slope, or in the autoregressive coefficients (e.g., in the cointegrating vectors). Similarly, the coefficient π_{11} is unique; hence, the relative difference in the magnitude of the variances in the mixture is the same for all coefficients. These constraints appear overly stringent. I do not see why one should expect the same probability of changes for all coefficients (or for changes in different components of the trend function of the noise function). I believe that substantial gains could be achieved by relaxing this constraint.

Concluding Comments

The main lesson that I have learned from having to study the issues behind the forecasting methodology proposed by Sims, to write these comments, is that I came to appreciate its qualities better. I expressed some divergence of opinion as to its Bayesian interpretation, but such issues are mainly ones of semantics and in no way question the fact that this forecasting model is well grounded and a very useful development in this line of research. My other comments are merely suggestions for possible extensions that could improve what is already a successful model.

While the model presented here appears relatively successful at providing unconditional predictions, it falls short when considering conditional predictions to analyze policy interventions and interpret the more structural aspects of the model. These issues can be analyzed only in a carefully identified system. Such identification issues have been studied in the context of VAR models with time-invariant parameters (see, e.g., Blanchard and Watson 1986; and Sims 1986) but are still open questions in the present, more general framework with time-varying parameters. Analyses pertaining to these identification issues indeed appear to be important avenues for future research.

References

Andrews, D. W. K. 1991. Exactly unbiased estimation of first order autoregressive/unit root models. Discussion Paper no. 975. New Haven, Conn.: Cowles Foundation.

Blanchard, O. J., and M. W. Watson. 1986. Are all business cycles alike? In *The American business cycle*, ed. R. J. Gordon. Chicago: University of Chicago Press.

Campbell, J. Y., and P. Perron. 1991. Pitfalls and opportunities: What macroeconomists should know about unit roots. *NBER Macroeconomics Annual*, 141–201.

Chen, C., and G. C. Tiao. 1990. Random level-shift time series models, ARIMA approximations, and level-shift detection. *Journal of Business and Economic Statistics* 8:83–98.

Fair, R. C. 1984. *Specification, estimation and analysis of macroeconometric models*. Cambridge, Mass.: Harvard University Press.

Gregory, A. W., and B. E. Hansen. 1992. Residual-based tests for cointegration in models with regime shifts. Working Paper no. 335. University of Rochester Center for Economic Research.

Hamilton, J. D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57:357–84.

Johansen, S. 1988. Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control* 12:231–54.

Kitagawa, G. 1987. Non-Gaussian state space modeling of nonstationary time series. *Journal of the American Statistical Association* 82:1032–41.

Martin, R. D., and A. E. Raftery. 1987. Comments: Robustness, computation and non-Euclidean models. *Journal of the American Statistical Association* 82:1044–50.

Park, J. 1992. E-M estimation of nonstationary state-space models with mixture Gaussian long-run errors: A structural change estimation. Princeton University. Mimeo.

Perron, P. 1988. Trends and random walks in macroeconomic time series: Further

evidence from a new approach. *Journal of Economic Dynamics and Control* 12:297–332.

———. 1989. The Great Crash, the oil price shock and the unit root hypothesis. *Econometrica* 57:1361–1401.

———. 1990. Testing for a unit root in a time series with a changing mean. *Journal of Business and Economic Statistics* 8:153–62.

Sims, C. A. 1986. Are forecasting models usable for policy analysis? *Quarterly Review of the Federal Reserve Bank of Minneapolis* 10 (1): 2–16.

———. 1988. Bayesian skepticism on unit root econometrics. *Journal of Economic Dynamics and Control* 12:463–74.

———. 1991. VAR econometrics: An update. Yale University. Mimeo.

Stock, J. H., and M. W. Watson. 1991. A probability model of the coincident economic indicators. In *Leading economic indicators: New approaches and forecasting records*, ed. K. Lahiri and G. H. Moore. Cambridge: Cambridge University Press.

Uhlig, H. 1991. What macroeconomists should know about unit roots as well: The Bayesian perspective. Princeton University. Mimeo.