Bayesian Methods for Dynamic Multivariate Models

Christopher A. Sims and Tao Zha

Federal Reserve Bank of Atlanta
Working Paper 96-13
October 1996

Abstract: If multivariate dynamic models are to be used to guide decision-making, it is important that it be possible to provide probability assessments of their results. Bayesian VAR models in the existing literature have not commonly (in fact, not at all as far as we know) been presented with error bands around forecasts or policy projections based on the posterior distribution. In this paper we show that it is possible to introduce prior information in both reduced form and structural VAR models without introducing substantial new computational burdens. With our approach, identified VAR analysis of large systems (e.g., 20-variable models) becomes possible.

JEL classification: C11, C53

The authors thank Eric Leeper for comments on an earlier draft. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions of substance to Christopher A. Sims, Department of Economics, Yale University, 37 Hillhouse Avenue, New Haven, Connecticut 06511, 203/432-6292, sims@econ.yale.edu; and Tao A. Zha, Research Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713, 404/521-8353, 404/521-8956 (fax), tao.a.zha@ atl.frb.org.

Questions regarding subscriptions to the Federal Reserve Bank of Atlanta working paper series should be addressed to the Public Affairs Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713, 404/521-8020, http://www.frbatlanta.org.

Bayesian Methods for Dynamic Multivariate Models

Introduction

If multivariate dynamic models are to be used to guide decision-making, it is important that it be possible to provide probability assessments of their results, for example to give error bands around forecasts or policy projections. In Sims and Zha [1995] we showed how to compute Bayesian error bands for impulse responses estimated from reduced form vector autoregressions (VAR's) and from identified VAR's. We also explained there the conceptual and practical difficulties surrounding attempts to produce classical confidence bands for impulse responses. However in that paper we considered only various types of "flat" prior.

But if we are to take seriously the results from such models, we are forced either to make artificially strong assumptions to reduce the number of parameters, or to follow Litterman [1986] in introducing Bayesian prior information. In this paper we show that it is possible to introduce prior information in natural ways, without introducing substantial new computational burdens. Our framework is developed for what are known as "identified VAR" models, but it includes as a special case reduced form models with the Litterman prior on its coefficients.

The developments we describe here are important for at least two reasons. The identified VAR literature has been limited for the most part to working with models of 6 to 8 variables, probably because sampling error makes results erratic in larger models under a flat prior. With our approach, identified VAR analysis of larger systems becomes possible. But even reduced form modeling under Litterman's prior has not before now been handled in an internally consistent way. The widely used method of constructing posterior distributions for VAR models

that is packaged with the RATS computer program is justified only for the case of priors that have the same form in every equation of the system. Usually it has been applied to models estimated "without" a prior – i.e. with a flat prior, which is trivially symmetric across equations. Litterman's prior differs across equations, because it treats "own lags" as different from other coefficients. Bayesian VAR models have therefore not commonly, in fact not at all as far as we know, been presented with error bands on forecasts or impulse responses based on the posterior distribution.

We consider linear multivariate models of the general form

$$A(L)y(t)+C=\varepsilon(t), \tag{1}$$

where $y(t)$ is an $m\times 1$ vector of observations, $A(L)$ is an $m\times m$ matrix polynomial of lag operator L with lag length p and non-negative powers, and C is a constant vector. We assume

$$\varepsilon(t)|\, y(s),\, s<t \;\sim\; N(0,\underset{m\times m}{I}) \,. \tag{2}$$

Though we work with this model, in which the only exogenous component is the constant vector, much of our discussion generalizes easily to more complicated sets of exogenous regressors. We assume $A(0)$ is non-singular so that (1) and (2) provide a complete description of the p.d.f. for the data $y(1), y(2),..., y(T)$ conditional on the initial observations $y(-p+1),..., y(0)$.

General Bayesian Framework: Identified Approach

The recent identified VAR models that aim at identifying monetary policy effects (e.g., Sims [1986], Gordon and Leeper [1994], Cushman and Zha [1995], Bernanke and Mihov [1996])

invoke economically interpretable restrictions on coefficients to make results interpretable. Such models work directly with the parameters in $A(L)$ from (1). The likelihood function is then

$$L\big(y(t),t=1,...,T|A(L)\big) \propto |A(0)|^T \exp\left[-\frac{1}{2}\sum_t \big(A(L)y(t)+C\big)'\big(A(L)y(t)+C\big)\right]. \tag{3}$$

Rewrite model (1) in matrix form:

$$\mathbf{Y}\mathbf{A}_0 - \mathbf{X}\mathbf{A}_+ = \mathbf{E}, \tag{4}$$

where $\mathbf{Y}$ is $T \times m$, $\mathbf{A}_0$ is $m \times m$, $\mathbf{X}$ is $T \times k$, $\mathbf{A}_+$ is $k \times m$, and $\mathbf{E}$ is $T \times m$. Note that $\mathbf{X}$ contains the lagged $\mathbf{Y}$'s and a column of 1's corresponding to the constant, T is the number of observations, m is the number of equations, and $k = mp+1$ is the number of coefficients corresponding to $\mathbf{X}$. Note that the arrangement of the elements in $\mathbf{A}_0$ is such that the columns in $\mathbf{A}_0$ correspond to the equations. That is to say, $\mathbf{A}_0 = A(0)'$.

Let

$$\mathbf{Z} = \begin{bmatrix}\mathbf{Y} & -\mathbf{X}\end{bmatrix}, \text{ and } \mathbf{A} = \begin{bmatrix}\mathbf{A}_0 \\ \mathbf{A}_+\end{bmatrix}. \tag{5}$$

We now introduce $\mathbf{a}$ as notation for $\mathbf{A}$ vectorized, i.e. the $m\cdot(k+m)\times 1$ vector formed by stacking the columns of $\mathbf{A}$, first column on top, and $\mathbf{a}_0$ and $\mathbf{a}_+$ correspondingly as notation for vectorized $\mathbf{A}_0$ and $\mathbf{A}_+$ respectively. Note that $\mathbf{a}_0$ and $\mathbf{a}_+$, though made up of elements of $\mathbf{a}$, do not arise from a simple partition of $\mathbf{a}$.

The conditional likelihood function (3) can now be expressed in compact form:

$$L(\mathbf{Y}|\mathbf{A}) \propto |\mathbf{A}_0|^T \exp\left[-0.5 trace(\mathbf{Z}\mathbf{A})'(\mathbf{Z}\mathbf{A})\right]$$

$$\propto |\mathbf{A}_0|^T \exp\left[-0.5\mathbf{a}'(I \otimes \mathbf{Z}'\mathbf{Z})\mathbf{a}\right] \tag{6}$$

Let us assume $\mathbf{a}$ has prior p.d.f.

$$\pi(\mathbf{a}) = \pi_0(\mathbf{a}_0)\varphi\left(\mathbf{a}_+ - \mu(\mathbf{a}_0); H(\mathbf{a}_0)\right), \tag{7}$$

where $\varphi(\cdot\,;\Sigma)$ is the standard normal p.d.f. with covariance matrix Σ. Of course one special case of (7) occurs when π is itself a normal p.d.f. in the full $\mathbf{a}$ vector. Combining (6) and (7), we arrive at the posterior density function of $\mathbf{a}$:

$$\begin{aligned} q(\mathbf{a}) \propto \pi_0(\mathbf{a}_0)|A(0)|^T |H(\mathbf{a}_0)|^{-1/2} \exp\Big[-0.5\big(&\mathbf{a}_0'(I \otimes \mathbf{Y}'\mathbf{Y})\mathbf{a}_0 + 2\mathbf{a}_+'(I \otimes \mathbf{X}'\mathbf{Y})\mathbf{a}_0 \\ &+ \mathbf{a}_+'(I \otimes \mathbf{X}'\mathbf{X})\mathbf{a}_+ + (\mathbf{a}_+ - \mu(\mathbf{a}_0))' H(\mathbf{a}_0)^{-1}(\mathbf{a}_+ - \mu(\mathbf{a}_0))\big)\Big] \end{aligned} \tag{8}$$

The posterior density (8) is non-standard in general, and the dimension of the parameter vector $\mathbf{a}$ is large even in relatively small systems of equations. A direct approach to analysis of the likelihood may therefore not be computationally feasible. However, the exponent in (8) is quadratic in $\mathbf{a}_+$ for fixed $\mathbf{a}_0$, meaning that the conditional distribution of $\mathbf{a}_+$ given $\mathbf{a}_0$ is Gaussian, making possible easy Monte Carlo sampling and analytic maximization or integration along the $\mathbf{a}_+$ dimension.

Symmetry

Though maximization or integration of (8) conditional on a fixed value of $\mathbf{a}_0$ is "only" a matter of linear algebra, the computations involved can be heavy because of their high dimensionality. The $\mathbf{a}_+$ vector is of order $m \cdot (mp+1)$, so finding its conditional posterior mean will require, at each value of $\mathbf{a}_0$, a least-squares calculation of that order. The calculation has the same form as that for a seemingly-unrelated-regressions (SUR) model. When there is no special structure, such computations are manageable for models with, say, $m = 6$ and $p = 4$, making the order of $\mathbf{a}_+$ 150, as might be realistic for a small quarterly model. But we have

applied these methods to models with 6 lags on 20 variables, and even tested them on models with 13 lags on 20 variables. For a 6-lag, 20-variable model $\mathbf{a}_+$ is of order 2440. With thirteen lags the order is 5220. Repeatedly solving least-squares problems of this order, using general algorithms, over many hundreds of iterations is impractical on widely available workstations.

On the other hand, because the calculation is of a SUR type, it has the usual property that it breaks into m separate least-squares calculations, each of dimension only $mp+1$, when the matrix of "regressors" is common across equations. It was this observation that led Litterman [1986] (and subsequent followers of his approach) to use single-equation methods on his reduced form model, even though the prior he proposed satisfies the conditions needed to give conditional likelihood the common-regressors form at best approximately. Highfield [1987] pointed out that by modifying Litterman's prior to make it symmetric across equations in the appropriate sense, one could make the full system posterior p.d.f. tractable. He was considering a reduced form model, i.e. a special case of ours in which $\mathbf{A}_0 = I$. In our more general framework we can also give the prior a form that makes the conditional posterior of $\mathbf{a}_+$ tractable, under conditions that are in some ways less restrictive than those Highfield required.

The demanding part of the calculations required for integrating or maximizing (8) with respect to $\mathbf{a}_+$ is a matrix decomposition of the coefficient on the term quadratic in $\mathbf{a}_+$. This coefficient can be read off from (8) as

$$(I \otimes \mathbf{X}'\mathbf{X}) + H(\mathbf{a}_0)^{-1} \ . \tag{9}$$

Clearly to preserve the Kronecker-product structure of the first term in this expression, we will require that

$$H(\mathbf{a}_0) = B \otimes G \,, \tag{10}$$

where B and G have the same order as the I and $\mathbf{X}'\mathbf{X}$ in (9). Further, either B must be a scalar multiple of I, or G must be a scalar multiple of $\mathbf{X}'\mathbf{X}$, because otherwise the Kronecker-product structure will be lost after the summation. Because $\mathbf{X}'\mathbf{X}$ depends on random variables generated by the model, it does not make sense to have our prior distribution's form depend on $\mathbf{X}'\mathbf{X}$. Thus preserving Kronecker-product structure requires that B be scalar, i.e. that beliefs about coefficients on lagged variables in structural equations have precision that is independent across equations. Since there is just a single G matrix, (10) requires also that the precision of beliefs about coefficients be the same in every equation.

Once B has been restricted to be scalar, though, the computational advantages of the strict Kronecker-product structure are no longer decisive. Suppose we have a distinct covariance matrix G_i for the prior on each equation's component of $\mathbf{a}_+$, but maintain independence across equations. Then (9) becomes

$$(I \otimes \mathbf{X}'\mathbf{X}) + diag(G_1,...,G_m) = diag(G_1 + \mathbf{X}'\mathbf{X},...,G_m + \mathbf{X}'\mathbf{X}) \,, \tag{11}$$

where we have introduced the notation

$$diag(G_1,...,G_m) = \begin{bmatrix} G_1 & 0 & \cdots & 0 \\ 0 & G_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & G_m \end{bmatrix} . \tag{12}$$

While a matrix decomposition of the right-hand side of (11) is not as easy as a decomposition of a Kronecker product, it is still far easier than a decomposition of a general $k \times k$ matrix, because it can be done one block at a time for the diagonal blocks. In our example of a 20-

variable, six-lag system, we are replacing a 2420×2420 decomposition with twenty 121×121 decompositions. Since these decompositions generally require computation time that is of cubic order in the size of the matrix, we have reduced the computations by a factor of 400.

It is interesting to contrast the situation in this model with what emerges from Highfield's consideration of SUR symmetry restrictions for Bayesian system estimation of reduced form VAR's. The straightforward approach to the reduced form leaves the covariance matrix of disturbances free, so that in place of the $I \otimes \mathbf{X}'\mathbf{X}$ first term in (9), we have a $\Sigma^{-1} \otimes \mathbf{X}'\mathbf{X}$ term, where Σ is the covariance matrix of the equation disturbances. This means that, to preserve a convenient system structure, prior beliefs must be treated as correlated across equations of the reduced form in the same pattern as Σ, an apparently unreasonable restriction. (We will see below that the restriction looks more plausible if we begin in a simultaneous equations framework.) The requirement that the covariance matrix of the prior be the same in every equation is also restrictive, and in this context, where the prior is directly on the reduced form, relaxing this requirement does greatly increase computational problems. In Litterman's approach, for example, the variances of coefficients on lags of the dependent variable in a reduced-form equation are larger than the variances of coefficients on other variables. This contradicts the requirement that the covariance matrix have the same structure in every equation.

Formulating a Prior Distribution

With the prior formulated as in (7), with a marginal p.d.f. on $\mathbf{a}_0$ multiplying a conditional p.d.f. for $\mathbf{a}_+ | \mathbf{a}_0$, our setup to this point has placed no restrictions on the conditional mean of $\mathbf{a}_+$. It restricts beliefs about $\mathbf{a}_+$ to be Gaussian and uncorrelated across equations conditional on $\mathbf{a}_0$,

but allows them to be correlated in different ways in different equations. This leaves many degrees of freedom in specifying a prior, making the use of substantive economic knowledge to form of a multivariate prior in these models a substantial task. In this section we suggest some approaches to the task.

A Base: The Random Walk Prior

The Litterman prior for a reduced form model expresses a belief that a random-walk model for each variable in the system is a reasonable "center" for beliefs about the behavior of the variables. Since this idea concerns behavior of the reduced form, it does not in itself restrict $\mathbf{A}_0$. It suggests that beliefs about the reduced form coefficient matrix

$$\mathbf{B} = \mathbf{A}_+\mathbf{A}_0^{-1} \tag{13}$$

should be centered on an identity matrix for the top m rows and zeros for the remaining rows. We make this notion concrete by making the conditional distribution for $\mathbf{A}_+$ Gaussian with mean of $\mathbf{A}_0$ in the first m rows and 0 in the remaining rows, or

$$E[\mathbf{A}_+|\mathbf{A}_0] = \begin{bmatrix} \mathbf{A}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{14}$$

As a starting point, we assume the prior conditional covariance matrix of the coefficients in $\mathbf{A}_+$ follows the same pattern that Litterman gave to the prior covariance matrix on reduced form coefficients. That is, we make the conditional prior independent across elements of $\mathbf{A}_+$ and with the conditional variance of the coefficient on lag ℓ of variable j in equation i given by

$$\frac{\lambda_1}{\sigma_j \ell^{\lambda_3}} . \tag{15}$$

The hyperparameter λ_1 controls what Litterman called overall tightness of beliefs around the random walk prior; λ_3 controls the rate at which prior variance shrinks with increasing lag length. The vector of parameters $\sigma_1,...,\sigma_m$ are scale factors, allowing for the fact that the units of measurement or scale of variation may not be uniform across variables. While in principle these should be chosen on the basis of a priori reasoning or knowledge, we have in practice followed Litterman in choosing these as the sample standard deviations of residuals from univariate autoregressive models fit to the individual series in the sample.

This specification differs from Litterman's in a few respects. There is no distinction here between the prior conditional variances on "own lags" versus "others" as there is in Litterman's framework. Because our model is a simultaneous equations model, there is no dependent variable in an equation, other than what might be set by an arbitrary normalization, so the "own" versus "other" distinction among variables is not possible. But note also that the unconditional prior for the top m rows in $\mathbf{A}_+$ will be affected by the prior on $\mathbf{A}_0$. In fact the unconditional prior variance of an element of the first m rows of $\mathbf{A}_+$ will be, because of (14), the sum of the prior variance of the corresponding element of $\mathbf{A}_0$ and the conditional variance specified in (15). Thus if our prior on $\mathbf{A}_0$ puts high probability on large coefficients on some particular variable j in structural equation i, then the prior probability on large coefficients on the corresponding variable j at the first lag is high as well.

Litterman's specification also has the scale factors entering as the ratio σ_i/σ_j, rather than only in the denominator as in (15). This reflects the fact that our specification normalizes the variances of disturbances in the structural equations to one.

The last row of $\mathbf{A}_+$ corresponds to the constant term. We give it a conditional prior mean of zero and a variance controlled by a separate hyperparameter λ_4. However it is not a good idea in practice to work with a prior in which beliefs about the constant term are uncorrelated with beliefs about the coefficients on lagged y's. Some of our suggestions below for modifying the prior via dummy observations are aimed at correcting this deficiency in the base setup.

The fact that this prior has a structure similar to Litterman's and can be similarly motivated should not obscure the fact that, because it is a prior on the conditional distribution of $\mathbf{A}_+|\mathbf{A}_0$ rather than on $\mathbf{B}|\mathbf{A}_0$, it entails different beliefs about the behavior of the data. In particular, as can be seen from (13), the prior described here makes beliefs about $\mathbf{B}$ correlated across equations in a way dependent on beliefs about $\mathbf{A}_0$, or equivalently about the covariance matrix of reduced form disturbances. Indeed in the special case where the prior covariance matrices G_i are the same across equations, the prior conditional distribution for $\mathbf{B}|\mathbf{A}_0$ is Gaussian with covariance matrix

$$\Sigma \otimes G \ , \tag{16}$$

which is exactly the form assumed by Highfield. Recall, though, that unlike Highfield we can also conveniently handle the case of differing G_i's, where there is no Kronecker product structure like (16) for the prior conditional covariance matrix of $\mathbf{B}$.

Dummy Observations, Unruly Trends

Litterman's work exploited the insight of Theil mixed estimation, that prior information in a regression model can be introduced in the form of extra "dummy" observations in the data matrix. A similar idea applies to the simultaneous equations framework we are considering here. For example, suppose we want to follow Litterman in starting with a prior centered on a reduced form implying the data series all follow random walks, correlated only through the correlation of innovations. Litterman, following an equation-by-equation approach to estimation, could implement his prior by adding to the data matrix used for estimating the i'th equation a set of $k-1$ dummy observations indexed by $j=1,...,m,\ \ell=1,...,p$, with data taking the values

$y_i(t)$	$y_r(t-s)$
$\begin{cases} \mu_1\mu_2^{\delta(i,j)}\sigma_i\,\ell^{\lambda_3} & i=j,\ \ell=1 \\ 0 & i\neq j \text{ or } \ell\neq 1 \end{cases}$	$\begin{cases} \mu_1\mu_2^{\delta(i,j)}\sigma_j\,\ell^{\lambda_3} & j=r,\ s=\ell \\ 0 & j\neq r \text{ or } s\neq \ell \end{cases}$

Here we have defined

$$\delta(i,j)=\begin{cases}0, & i\neq j\\ 1, & i=j\end{cases} \tag{17}$$

We are also introducing a convention that scale factors for variances in the prior covariance matrix are λ's, while scale factors on dummy observations are μ's. When the same distribution can be formulated either directly with a prior covariance matrix or indirectly via dummy observations, the similarly numbered λ's and μ's correspond, with $\lambda_i=1/\mu_i^2$. The i'th equation's dummy observations can be written as

$$\mathbf{Y}_{id}\mathbf{A}_{0\bullet i}=\mathbf{X}_{id}\mathbf{A}_{+\bullet i}+\mathbf{E}_{\bullet i}\ , \tag{18}$$

where a "$\bullet i$" subscript on a matrix refers to the i'th column of the matrix.

In our approach, where all equations are estimated jointly, the fact that these "dummy observations" are equation-specific means that they are not algebraically equivalent to adding rows to the data matrix. One might nonetheless introduce them into the exponent in the posterior p.d.f. (8) as

$$\mathbf{a}_0' \cdot diag\left(\left\{\mathbf{Y}_{id}'\mathbf{Y}_{id}\right\}_{i=1}^{m}\right)\cdot \mathbf{a}_0 + 2\mathbf{a}_0' \cdot diag\left(\left\{\mathbf{Y}_{id}'\mathbf{X}_{id}\right\}_{i=1}^{m}\right)\cdot \mathbf{a}_+ + \mathbf{a}_+' \cdot diag\left(\left\{\mathbf{X}_{id}'\mathbf{X}_{id}\right\}_{i=1}^{m}\right)\cdot \mathbf{a}_+ \quad . \qquad (19)$$

These terms do not introduce any complications in the numerical analysis, because they preserve the block diagonal structure of the coefficient matrix for the term quadratic in $\mathbf{a}_+$. In fact, terms of this form can be used to implement exactly the conditional prior for $\mathbf{A}_+|\mathbf{A}_0$ described in the preceding section. To implement that prior we would want to set $\mu_2 = 1$ and $\mu_1 = 1/\sqrt{\lambda_1}$ in the formulas in the table above.

In work following Litterman's, modifications of his prior have been introduced that improve forecasting performance and take the form of true, system-wide dummy observations. The "sums of coefficients" component of a prior, introduced in Doan, Litterman, and Sims [1984], expresses a belief that when the average of lagged values of a variable is at some level $\bar{y}_i$, that same value $\bar{y}_i$ is likely to be a good forecast of $y_i(t)$. It also implies that knowing the average of lagged values of variable j does not help in predicting a variable $i \neq j$. In a system of m equations it introduces m observations, indexed by j, of the form

$y_i(t)$	$y_r(t-s)$
$\begin{cases} \mu_5 & i = j \\ 0 & i \neq j \end{cases}$	$\begin{cases} \mu_5 & j = r, \text{ all } s \\ 0 & j \neq r \end{cases}$

In these dummy observations, the last column of the data matrix, corresponding to the constant term, is set to zero. These dummy observations introduce correlation among coefficients on a given variable in a given equation. When $\mu_5 \to \infty$, the model tends to a form that can be expressed entirely in terms of differenced data. In such a limiting form, there are as many unit roots as variables and there is no cointegration.

The "dummy initial observation" component of a prior, introduced by Sims [1993], introduces a single dummy observation in which, up to a scaling factor, all values of all variables are set equal to the corresponding averages of initial conditions, and the last column of the data matrix is set at its usual value of 1. We designate the scale factor for this dummy observation as μ_6. This type of dummy observation reflects a belief that when lagged values of y_i have averaged $\bar{y}_i$, that same value $\bar{y}_i$ should be a good forecast, but without any implication that there are no cross effects among variables or that the constant term is small. This kind of dummy observation introduces correlations in prior beliefs about all coefficients in a given equation. As $\mu_6 \to \infty$, the model tends to a form in which either all variables are stationary with means equal to the sample averages of the initial conditions, or there are unit root components without drift (linear trend) terms. The number of unit root terms in the limit as $\mu_6 \to \infty$ is indeterminate, so cointegrated models are not ruled out in this limit.

Both of these types of dummy observation are symmetric across equations, so that they can be introduced as extra rows in the data matrix, making them easy to handle efficiently in computation.

These latter two types of true dummy observations, taken together, favor unit roots and cointegration, which fits the beliefs reflected in the practices of many applied macroeconomists. More importantly, they have been found to improve forecasts in a variety of contexts with economic time series.

Some insight into why this should be so is provided in Sims [1992], which shows that without such elements in the prior, fitted multivariate time series models tend to imply that an unreasonably large share of the sample period variation in the data is accounted for by deterministic components. That is, if we construct from the estimated coefficients, treating them as non-random, the vector of time series $E[y(t)|y(s), s \le 0]$, t=1,...,T, we find that they show substantial variation, while remaining close to $y(t)$ itself, even for large values of t. This bias toward attributing unreasonable amounts of variation to deterministic components is the other side of the well-known bias toward stationarity of least-squares estimates of dynamic autoregressions.

We do not have a clear theoretical explanation for why estimated multivariate time series models should show stronger bias of this type than do lower dimensional models, but it seems that they do. One can certainly see how it is possible. A reduced-form system in m variables with p lags can generally fit without error an arbitrary collection of time series that are all polynomials in t of order $mp-1$. The collection of time series and their first $m-1$ lags will in general all be linearly independent and all will by construction lie in the space spanned by the 0'th through $mp-1$'th power of t. They therefore form a basis for the space of $mp-1$'th order

polynomials in t. Thus some linear combination of them exactly matches the dependent variable, which in each equation of the reduced form system is itself by assumption such a polynomial.

This result means that a least-squares algorithm, attempting to fit, say, 6 time series with a 5'th order VAR, always has the option of taking a form in which $E[y(t)|y(s), s \le 0]$, t=1,...,T is a freely chosen set of 29'th order polynomials in t. Such high order polynomials are likely to be able to fit many economic time series quite well, while still being implausible for out-of-sample projections. Of course, this argument is only heuristic, because it has not shown that VAR coefficients that exactly fit high-order polynomial approximations to the data series will provide a good fit to the data series themselves. Nonetheless, there seems to be cause for concern about overfitting of low-frequency deterministic components both from a theoretical point of view and, as shown in Sims [1992], from an applied perspective.

A Prior on $\mathbf{A}_0$

In many applications the prior on $\mathbf{A}_0$ will reflect the substantive knowledge that makes it possible to distinguish the equations as behavioral mechanisms – i.e., the identifying restrictions. Since this paper is concerned mainly with explaining the econometric technique, we do not here discuss how to come up with identifying restrictions. We should note, though, that in this approach it is often desirable to aim for distinct behavioral interpretations only of blocks of equations, not the complete list of individual equations. Within a block of equations that are not separately identified, we can simultaneously make coefficients unique and the disturbance matrix diagonal by a triangular normalization – we impose zero restrictions on an otherwise unrestricted

square block of coefficients within the block of equations so as to force it to take on a triangular form.

A reduced form model can be estimated within the framework of this section by taking $\mathbf{A}_0$ to be triangular, as a normalization, and imposing no other prior restrictions on it. The prior on $\mathbf{A}_0$ is then equivalent to a prior on the reduced form innovation covariance matrix Σ.

Examples

We display results for two cases: One matching the structure laid out in the preceding sections, with a simple exactly identified parameterization of $\mathbf{A}_0$ and a Gaussian prior on its elements; one matching the usual interpretation of Litterman's prior, in which it is independent across elements of $\mathbf{B}|\mathbf{A}_0$ rather than $\mathbf{A}_+|\mathbf{A}_0$, and using a non-Gaussian "flat" prior on $\mathbf{A}_0$. We show that the former results in much more convenient calculations, that nonetheless give quite reasonable results. We also show that some apparently plausible numerical shortcuts for the latter example, that have appeared in the literature, can have substantial effects on results.

The variables in both models are quarterly data on: the 3-month T-bill rate (R), money stock (M1), real GNP (y, \$1982), GNP deflator (P), the unemployment rate (U), and gross private domestic fixed investment (I). Money stock, real GNP, GNP deflator, and fixed investment are all in logarithms, and the sample period is 48:1-82:4. Also in both models we include 4 lags, i.e. set $p = 4$.

Prior Independence Across Structural Equations

We apply the "identified" Bayesian method to a reduced form model. As a normalization, we restrict $\mathbf{A}_0$ to be upper triangular. We make the prior on $\mathbf{A}_0$ Gaussian, with the prior standard

deviation on each coefficient in the j'th column of $\mathbf{A}_0$ set to $\lambda_0 / \hat{\sigma}_j$, where λ_0 is a new hyperparameter. The prior on $\mathbf{A}_+|\mathbf{A}_0$ is of the same form we have described above, including both types of dummy observations.

We generated draws from the posterior distribution of the data for the period after 1982:4, the end of the sample period. In the prior, we set the weights μ_5 and μ_6 on the two types of dummy observations to 1, $\lambda_0 = 1$, $\lambda_1 = 0.2$, and $\lambda_2 = \lambda_3 = \lambda_4 = 1$. Figure 1 displays the results, based on 5000 MC draws. The solid lines shown are the actual data series. The central dotted line, made up of alternating long and short dashes, is the posterior mean, and the two outer dotted lines represent 16th and 84th percentiles, so that the bands contain about the same amount of probability (68%) as one-standard-error bands. The bands are calculated pointwise, so that the posterior probability of the future data being in the band is 68% at each forecast horizon individually, not for the band as a whole.

Our numerical procedure was first to maximize the marginal posterior on $\mathbf{a}_0$, then to use importance sampling. That is, we drew values of $\mathbf{a}_0$ from the multivariate t-distribution with 9 degrees of freedom centered on the posterior mode and with covariance matrix given by the Hessian at this mode. To make the results reflect the true posterior distribution, we weighted $\mathbf{a}_0$ draws by the ratio of the true posterior p.d.f. to the p.d.f. of the multivariate t from which we were drawing. For each draw of $\mathbf{a}_0$ generated this way, we generated an associated $\mathbf{a}_+$ by drawing from the conditional normal posterior on $\mathbf{a}_+$. This of course involved solving a large least squares problem at each draw.

This is a difficult period for VAR models to forecast, particularly for prices. The bands and forecasts nonetheless look reasonable. The actual data lie in or close to the 68% posterior band except for prices. The price forecast predicts substantially more inflation than actually occurred, and gives very low probability to actual values as far from the forecast as actually occurred. The tendency of fixed-coefficient Bayesian VAR's to do badly in forecasting prices in the 80's was a main motivation for the extensions to the BVAR framework introduced in Sims [1993].

Prior Independence Across Reduced-Form Equations

We now consider a different type of prior on the same model, aiming to match the usual interpretation of Litterman's prior on the reduced form. We keep the prior conditional on $\mathbf{A}_0$ in the same form as in sections 0 and 0, but interpret it now as applying to $\mathbf{B}|\mathbf{A}_0$ rather than to $\mathbf{A}_+|\mathbf{A}_0$. Also, to keep it in line with common usage of this sort of prior, we make $\lambda_2 = .3$, so that coefficients representing cross-variable effects are taken as a priori likely to be smaller, and to maintain comparability with Litterman's original work we drop the two kinds of true dummy observations, i.e. set $\mu_5 = \mu_6 = 0$. The joint posterior on $\mathbf{A}_0$ and $\mathbf{A}_+$ is still in the form (8), but with $\mu(\mathbf{a}_0)$ given by (14) and with

$$H(\mathbf{a}_0) = (\mathbf{A}_0' \otimes I) \cdot diag(\{G_i\}) \cdot (\mathbf{A}_0 \otimes I) \ . \tag{20}$$

Note that in going back and forth between a prior using the $\mathbf{A}_0, \mathbf{B}$ parameterization and one using the $\mathbf{A}_0, \mathbf{A}_+$ parameterization, the Jacobian, $|\partial \mathbf{B} / \partial \mathbf{A}_+| = |\mathbf{A}_0|^{-k}$, must be taken into account. We in fact use $|\mathbf{A}_0|^k$ as an improper prior for $\mathbf{A}_0$ here when writing the prior in terms of $\mathbf{B}$, so that when transformed to the $\mathbf{A}_0, \mathbf{A}_+$ parameter space, the prior on $\mathbf{A}_0$ is flat.

Clearly with the conditional covariance matrix in the form of (20), the outcome of the addition in (9) is not a matrix with any standard special structure we can exploit in a matrix decomposition algorithm. Nonetheless, the models we consider here are small enough, with $mp = 24$ and $k = m \cdot (mp + 1) = 150$, that direct manipulation of the 150×150 conditional covariance matrix is computationally feasible. We include this case in part to display the nature of the gains in computational convenience from using the formulation in the preceding section.

Figure 2 shows error bands computed with this prior. Besides the posterior mean and the 68% band shown in Figure 1, Figure 2 also shows (as a dashed line with equal-length long dashes) the forecast as Litterman originally constructed it, from single-equation estimates, ignoring randomness in the coefficients conditional on the data. It is apparent that the prior differs from that underlying Figure 1 and that this affects results. Particularly noticeable is the shift in the location of the unemployment forecast and its error band, with this latter estimate showing less decline in unemployment and with the actual path of unemployment largely outside the 68% band. The Figure 1 forecast is also more optimistic for output. Note that this was a period in which Litterman's model outperformed commercial forecasts by making forecasts for output that were much more optimistic than those of most commercial forecasters. The Figure 1 forecasts, perhaps because they embody prior belief in correlation of coefficients in reduced form equations related to correlation of innovations, make the optimism on unemployment match the optimism on output, while this is less true of the Figure 2 forecasts.

We do not mean by this comparison to imply that one of these priors is clearly better than the other as a standard for use in forecasting models. The graphs do suggest, though, that the priors

on $\mathbf{A}_+|\mathbf{A}_0$ produce reasonable estimates, comparable to those from a prior on $\mathbf{B}|\mathbf{A}_0$. But the $\mathbf{A}_+|\mathbf{A}_0$ posteriors are far more convenient for computation. The computing time for obtaining the peak of the marginal posterior on $\mathbf{a}_0$ for Figure 2, using our own code for optimization, written in Matlab,[1] is about 2.8 hours on a Pentium/120 machine, and additional computation for generating weighted Monte Carlo draws takes about 16 minutes per 1,000 draws. In contrast, the peak of the marginal posterior on $\mathbf{a}_0$ is obtained within 1 minute for Figure 1, and thereafter 1000 Monte Carlo draws take about 4 minutes. These ratios of times would become greater with larger systems. As noted above, we have used these methods with $\mathbf{A}_+|\mathbf{A}_0$ priors on overidentified VAR's, in which the $\mathbf{A}_0$ matrix is restricted, with up to 20 variables and 6 lags. Handling models of this size with a $\mathbf{B}|\mathbf{A}_0$ prior would be prohibitively time-consuming.

Conclusion

It is feasible to use Bayesian prior information in reduced form and structural VAR models, and to do so in a way that keeps computation growing only with the square of system size. This should make it possible to use larger VAR systems in forecasting and policy analysis, where a Bayesian approach is essential to avoid high mean-squared error.

[1] The Matlab code used for the maximization is available via the web at http://www.econ.yale.edu/~sims or directly via ftp at http://ftp.econ.yale.edu/pub/sims. It is more robust to deterioration of numerical accuracy and to certain kinds of one-dimensional discontinuities than are most such programs.

References

Bernanke, B.S. and I. Mihov, 1996. "Measuring Monetary Policy", *manuscript* (Princeton University).

Cushman, David O. and Tao Zha, 1995. "Identifying Monetary Policy in a Small Open Economy Under Flexible Exchange Rates," *Federal Reserve Bank of Atlanta Working Paper 95-7.*

Doan, Thomas, Robert Litterman, and Christopher A. Sims, 1984. "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, Vol.3, 1-100.

Gordon, D.B. and Eric. M. Leeper, 1994. "The Dynamic Impacts of Monetary Policy: An Exercise in Tentative Identification," *Journal of Political Economy 102*, 1228-1247.

Highfield, Richard A., 1987. "Forecasting with Bayesian State Space Models," Ph.D. dissertation, Graduate School of Business, University of Chicago.

Litterman, R.B., 1986. "Forecasting With Bayesian Vector Autoregressions -- Five Years of Experience," *Journal of Business & Economic Statistics 4*, 25-38.

Christopher A. Sims, 1986. "Are Forecasting Models Usable for Policy Analysis," *Quarterly Review* of the Federal Reserve Bank of Minneapolis, Winter.

_____, 1992. "Bayesian Inference for Multivariate Time Series with Trend," presented at the American Statistical Association meetings and available at http://www.econ.yale.edu/~sims/ or ftp://ftp.econ.yale.edu/pub/sims.

_____, 1993. "A Nine-Variable Probabilistic Macroeconomic Forecasting Model," in J.H. Stock and M.W. Watson (eds.), *Business Cycles, Indicators and Forecasting*, University of Chicago Press for the NBER, 179-204.

Sims, C.A. and T. Zha, 1995. "Error Bands for Impulse Responses," *Federal Reserve Bank of Atlanta Working Paper 95-6.*

Figure 1

82:4 Forecasts: Prior independence of orthogonalized equations

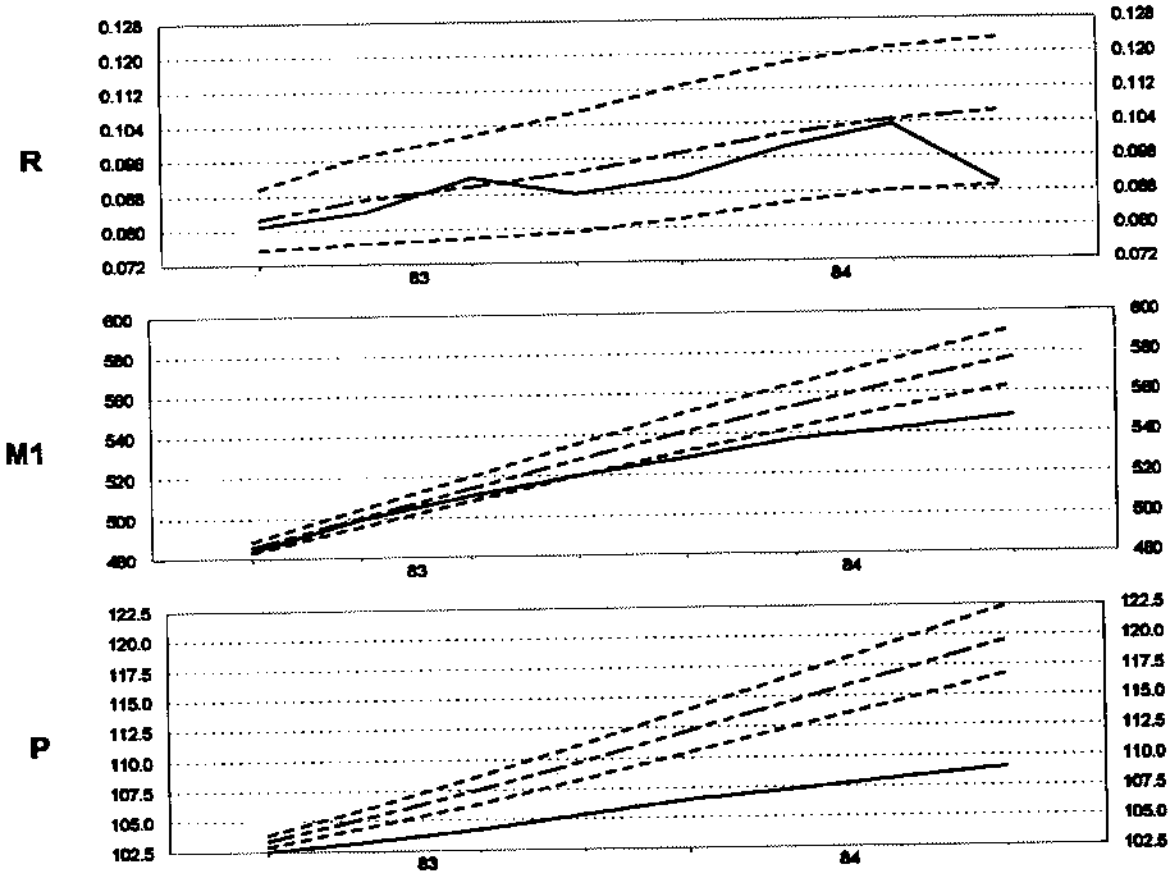

Figure 1 continued

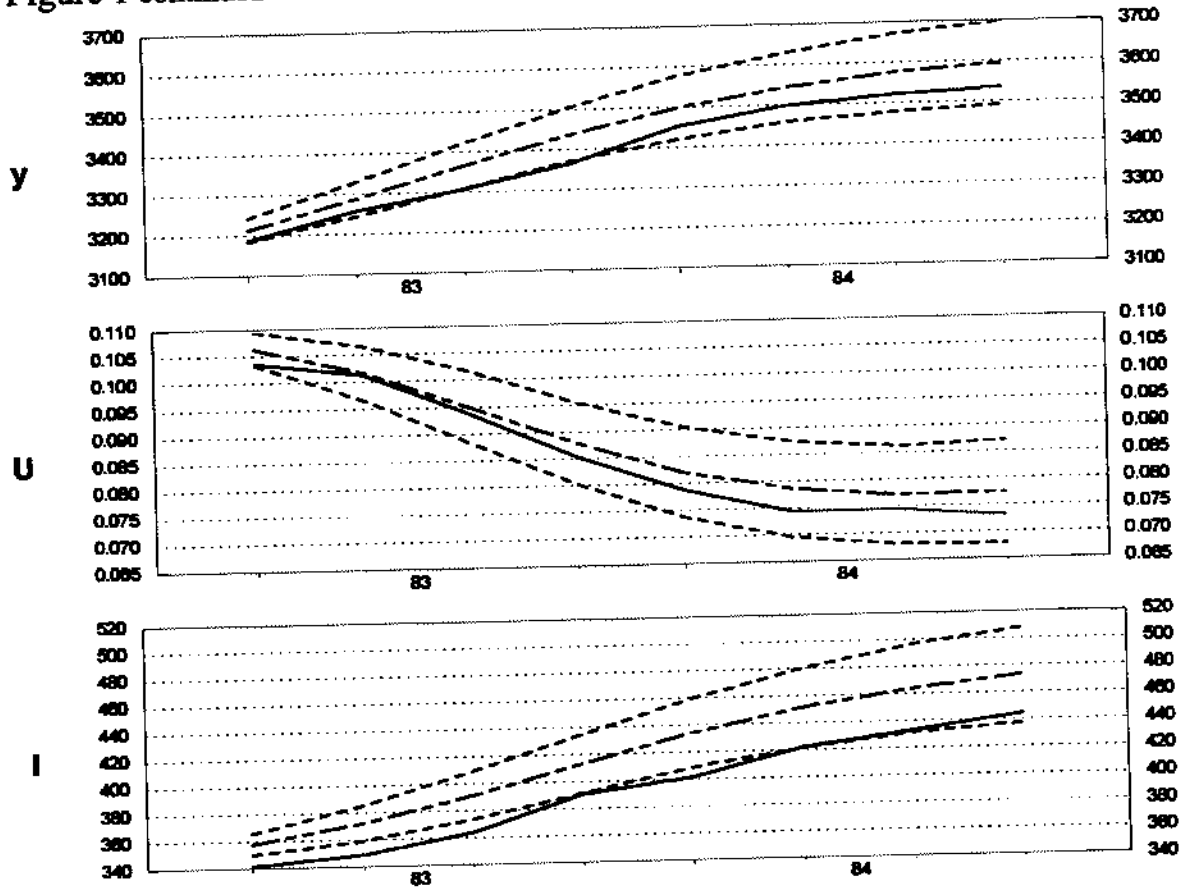

Figure 2

82:4 Forecasts: Prior independence of reduced-form equations

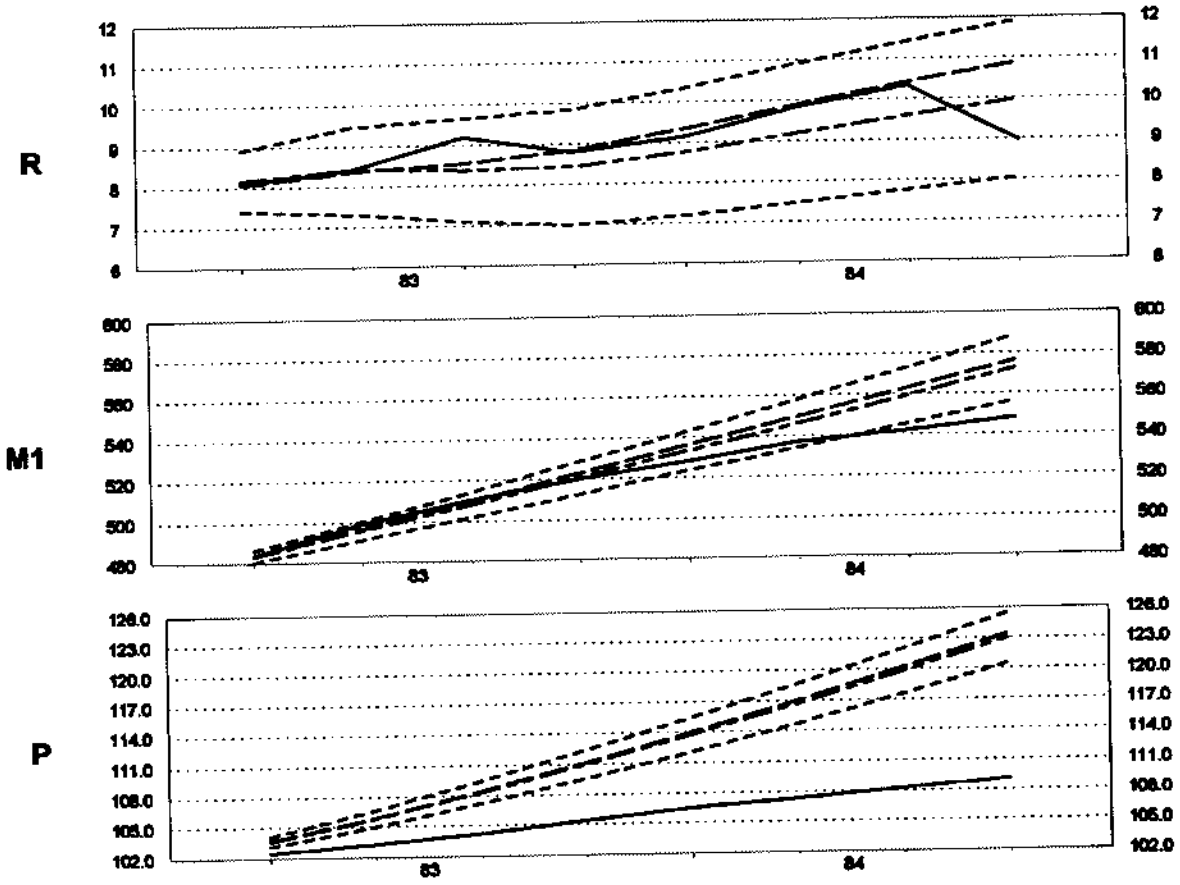

Figure 2 continued