CONJUGATE DUMMY OBSERVATION PRIORS FOR VAR'S

1. THE GENERAL IDEA

As is documented elsewhere Sims (revised 1996, 2000), there is a strong tendency for estimated time series models, when the estimates are based on a flat prior and on likelihood conditional on initial observations, to attribute unreasonable explanatory power to initial conditions. Since classical estimation methods (like OLS in simple unrestricted VAR's) are close to flat-prior Bayesian methods, they have this same tendency. The pathology is reflected, in classical distribution theory, in the small-sample bias toward stationarity in estimated models.

The problem with the flat-prior estimates is that in these models a flat prior implies that we give considerable credence to the possibility that in our sample initial conditions are very far from the model's steady state. If this were true, it would imply that the data contain an "initial transient" that allows the model to be estimated with much higher precision in the sample at hand than would be the case in a typical sample taken from data after the end of the current sample period. While this may be a realistic possibility for some applications, in most economic applications it is not.

One solution to this problem is to use the unconditional likelihood. That is, derive the mapping from the model's parameters to the unconditional distribution of the initial conditions, and combine that pdf with the usual likelihood function in doing inference. This approach requires the use of iterative nonlinear optimization methods, but it is quite feasible. The main disadvantage of this approach is that, because there is no unconditional distribution of the initial conditions if the model implies non-stationarity, the approach applies only when we are certain in advance that the model should imply stationarity, and this is seldom the case in economic applications.

Another approach is to introduce a prior that captures our belief that it is implausible that initial transients explain a large part of observed long-run variation, which is what we explain how to do here.

2. MECHANICS

The basic idea of dummy observation priors for VAR's is the same as that for dummy observation priors in ordinary regression models. Intuitively, one adds extra "data" to the sample that express prior beliefs about the parameters. The prior takes the form of the likelihood function for the dummy observations. The difference from single-equation regression models is that the added observations are used in all the equations of the system, not just in one.

Circumstances can arise where one has beliefs about parameters in some equations but not others, or has different beliefs about different equations. In fact, the "Minnesota Prior" that has been widely used in VAR forecasting models and is automated in the

RATS program has this character. It might seem we could handle such a prior just by creating separate sets of dummy observation for each equation, and indeed this is the way RATS and most of the research in the literature that uses priors has implemented them. However, once the dummy observations differ across equations, the seemingly unrelated regressions algebra that justifies equation-by-equation estimation no longer applies. So to maintain internal consistency in likelihood calculations, it is best to stick to pure system-wide dummy observations. Such dummy observations amount to stating prior beliefs about how the whole system will behave.

In all our discussion below we will assume a model in the form

$$(1) \qquad\qquad y(t) = c + \sum_{s=1}^{k} B_s y(t-s) + \varepsilon(t)$$

with data for $t = 1, \ldots, T$ and $\varepsilon(t) \mid \{y(t-s), s \geq 1; B, c, \Sigma\} \sim N(0, \Sigma)$.

2.1. **The one-unit-root prior.** The type of prior we describe here is called in the published literature a "dummy initial observation" prior. This is not very descriptive however, so we'll call it the one-unit-root prior. We introduce data for the artificial date t^* in which

$$\underset{n \times 1}{y(t^*)} = y(t^*-1) = \ldots = y(t^*-k) = \underset{n \times 1}{\bar{y}} \, \lambda$$

and the vector of 1's that corresponds to the constant term in the data matrix is set to λ in the t^* observation. The vector $\bar{y}$ is usually set to the sample mean of the initial conditions, i.e.

$$(2) \qquad\qquad \bar{y} = \frac{1}{k} \sum_{s=1}^{k} y(1-k) \, .$$

If we write out the equation for this observation, we get

$$(3) \qquad\qquad \bar{y}\lambda = \left(\sum_{s=1}^{k} B_s \right) \bar{y}\lambda + c\lambda + \varepsilon(t^*) \, .$$

If $I - B(1)$ is full rank, this equation can be rearranged to read

$$(4) \qquad\qquad \bar{y} = \left(I - B(1) \right)^{-1} c + \left(I - B(1) \right)^{-1} \varepsilon(t) \lambda^{-1} \, .$$

It might appear, then, that the dummy observation is equivalent to asserting an unconditional pdf for $\bar{y}$, i.e.

$$(5) \qquad \bar{y} \mid \{B, c, \Sigma\} \sim N \left(\left(I - B(1) \right)^{-1} c, \lambda^{-2} \left(I - B(1) \right)^{-1} \Sigma \left(I - B(1) \right)'^{-1} \right) \, .$$

In the case where $n = k = 1$, so $\bar{y} = y(0)$, this would simply be an unconditional pdf for $y(0)$. Otherwise, it is an unconditional pdf for a particular set of linear combination of elements of the initial conditions $\{y(0), \ldots, y(-k+1)\}$, so amounts to an improper unconditional pdf for the initial conditions. However, the dummy observation does not actually add to the log likelihood the same terms that we would add by using (5) as an unconditional pdf for the initial conditions. The dummy observation adds the terms

$$(6) \qquad\qquad -\frac{1}{2} \log |\Sigma| - \frac{\lambda^2}{2} \left((I - B(1))\bar{y} - c \right)' \Sigma^{-1} \left((I - B(1))\bar{y} - c \right) \, .$$

Using (5) as an unconditional distribution for $\bar{y}$ would add the same terms, but in addition would add $\log \lambda + \log |I - B(1)|$. The extra term in λ has no effect on the posterior for c, B and Σ, but the term in $B(1)$ has an important effect. In using the dummy observation, then, we are acting as if we have an unconditional distribution for $\bar{y}$ given by (5), and at the same time have an improper prior on B proportional to $|I - B(1)|^{-1}$.

The reason for calling this dummy observation a "single-unit root" prior is that $|I - B(1)| = 0$ is exactly the condition that there be at least one unit root in the system. Thus the prior is centered on a part of the parameter space where either $c = 0$ and the system contains a unit root with $\bar{y}$ as its eigenvector, or $c \neq 0$, y is stationary, and $y(0)$ is close to the model's implied population mean.

If one wants the prior to lean more sharply toward the presence of a unit root, one can introduce separate dummy observations in which all the y's are zero and the element of the data matrix to which c applies is non-zero. In conjunction with the single-unit-root prior as we have set it out, this forces c closer to zero and thus beliefs closer to a unit-root case. One could also add dummy observations that have the same form as the single-unit-root dummy observation, except that the column of the data matrix corresponding to c is set to zero. This type of dummy observation used alone would center on the region where there is at least one unit root and would not give a priori plausibility to any stationary models. If used in isolation, though, it puts no limits on c, and thus could allow estimates in which, despite the presence of a unit root, there are still strong initial transients.

3. No-cointegration dummy observations

Cointegration is the situation where a model has some stable roots and some unit roots. If $I - B(1)$ has m unit eigenvalues (with y $n \times 1$), none of which are repeated[1], then there are $n - m$ linear combinations $z_\gamma(t) = \gamma y(t)$ that are stationary, even though it may well be that every element of the y vector is non-stationary. The single-unit-root prior is consistent with the presence of cointegration, since it favors only the existence of some unit root.

A set of n dummy observations that centers on unit-root behavior separately for each element of y can be formed by taking as the j'th dummy observation one in which, at the artificial date t_j^*,

$$(7) \qquad y(t_j^*) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \bar{y}_j \lambda \\ 0 \\ \vdots \\ 0 \end{bmatrix} = y(t_j^* - 1) = \ldots = y(t_j^* - k) \, ,$$

_____

[1]In the sense that none gives rise to off-diagonal entries in the Jordan decomposition. The literature on cointegration has tended to focus on this case, even though it is hard to see why we should be confident a priori that unit roots will not repeat.

with the $\bar{y}_j$ in the bracketed vector occurring in the j'th position, and the column of the data matrix corresponding to c set to 0. This dummy observation then reads as, for the j'th equation

$$\lambda \bar{y}_j = \sum_s B_{jjs} \bar{y}_j \lambda + \varepsilon_j(t_j^*) \tag{8}$$

and for $k \neq j$,

$$0 = \sum_s B_{kjs} \bar{y}_j + \varepsilon_k(t_j^*). \tag{9}$$

It is easy to see that this prior favors B's whose off-diagonal elements are small, and also B's that are close to putting a unit root in the $B_{jj}(L)$ polynomial for each j.

This prior works to suppress initial transients, but it does much more than that. It may be useful as an approximate representation of widely shared prior beliefs (that unit roots are present, that cross-variable relations are weak) but it should not be justified solely by appeal to the idea that initial transients are implausible.

This kind of dummy observation also puts no limits on c. Thus if one wants to down-weight versions of the model in which there are unit roots and $c_j^2 >> \Sigma_{jj}$, i.e. models in which deterministic trend components are much more important than the error term, then one should combine these no-cointegration dummy observations with dummy observations that favor $c = 0$.

Note that dummy observation priors can be combined. It has been common in applied work to apply both single-unit-root and no-cointegration dummy observations.

4. THE MINNESOTA PRIOR

We are working back in historical time. The use of no-cointegration priors in applied work preceded the use of single-unit-root priors, but preceding them both was the Minnesota Prior. The Minnesota Prior postulates a separate set of dummy observations for each equation, implementing a prior that specifies independent distributions for all coefficients, with the coefficient B_{jks} having mean 0 unless $j = k$ and $s = 1$, in which case the mean is 1, and variance $\pi_1^{-1} \pi_2^{1-\delta(j,k)} s^{-\pi_3}$, where $\delta(j,k)$ is a function equal to 1 if $j = k$ and 0 otherwise. With $\pi_2 < 1$, this prior specifies that coefficients on own lags $(B_{jj}(s), s = 1, \ldots, k)$ are likely to be larger in absolute value than coefficients on other variables, and with $\pi_3 > 0$, it specifies that coefficients on more distant lags are likely to be smaller.

This prior has usually been implemented, inconsistently, via separate OLS estimates using a distinct set of dummy observations on each equation. As we have already pointed out, the usual ML or posterior mean interpretation of equation-by-equation OLS is not correct when the data matrix (dummy observations or otherwise) is different across equations. If we omit the π_2 term (that is, set it to 1), then own lag coefficients and cross-variable coefficients are treated symmetrically, and the prior can be implemented with system-wide dummy observations. If t_{ks}^* is the artificial date for the dummy observation applying to the s'th lag of the k'th variable, one of these dummy observations has the

form

$$(10) \qquad y(t_{k1}^*) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \pi_1\sigma_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} = y(t_{k1}^* - 1), \qquad y(t_{ks}^* - s) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \pi_1 s^{\pi_3}\sigma_k \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$(11) \qquad y(t_{ks}^* - v) = 0 \text{ for } v \neq s > 1 \text{ and for } v > s = 1.$$

The parameter σ_k is a measure of the degree of variability in the k'th variable. It has most commonly, again somewhat inconsistently, been set by estimating a low-order univariate AR regression for each variable and taking σ_k to be the standard deviation of the residual of that regression. It could also be set as the sample standard deviation of the initial conditions for each variable.

In a recent paper Sims and Zha (1998) a new approach to variable-by-variable setting of the prior has been proposed that captures the spirit of the Minnesota Prior without the inconsistencies and with nearly the same degree of convenience. The approach of that paper also generalizes to the case of structural VAR's (models with restrictions on Σ), which the Minnesota Prior does not. In this course, we will not have time to go into the details of this new approach.

5. EXERCISE DUE FRIDAY, 12/17

Repeat the VAR estimation and testing that you carried out for the last exercise, this time with a single-unit-root prior with $\lambda = 1$ on each system you estimate. Compare the results to those you obtained originally. For each estimated system, compute and plot the actual time series $y(t), t = 1, \ldots, T$ and the projections from initial conditions, $E_0[y(t)], t = 1, \ldots, T$. Was there a problem with initial transients to start with? Did the single-unit-root prior "fix" it?

REFERENCES

SIMS, C. A. (2000): "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples," *Journal of Econometrics*, 95(2), 443–462, http://www.princeton.edu/~sims/.

——— (revised 1996): "Inference for Multivariate Time Series with Trend," Discussion paper, presented at the 1992 American Statistical Association Meetings, http://sims.princeton.edu/yftp/trends/ASAPAPER.pdf.

SIMS, C. A., AND T. ZHA (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*.