

## Trend and Unit Root Anomalies; Dummy Observation Priors\*

### 1. WHAT'S THE PROBLEM?

We have already discussed the “Helicopter Tour” model, in which

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (1)$$

and we observed that with the  $\varepsilon_t$ 's i.i.d. normal, conditional on initial conditions, and with a flat prior,  $\rho | \hat{\rho}$  has a symmetric distribution about  $\hat{\rho}$ , while at the same time  $\hat{\rho} | \rho$  has an asymmetric distribution about  $\rho$ . We argued there that the classical bias in  $\hat{\rho} | \rho$  does not need to be “corrected”, unless there is good reason to use a non-flat prior on  $\rho$ . In this sense, the classical bias in the OLS estimate of  $\rho$  is not a problem.

However, in models with many variables, or with constants and trend terms, there is a problem. We can see this in a variety of ways. Several examples that illustrate the point in different ways follow.

**Univariate AR with constant—Which flat prior?:** Consider

$$y_t = c + \rho y_{t-1} + \varepsilon_t . \quad (2)$$

If  $|\rho| < 1$ , this model implies, if it has been in effect a long time, that

$$y_t = \frac{c}{1-\rho} + \frac{\varepsilon_t}{1-\rho L} . \quad (3)$$

If we let  $\mu = c/(1-\rho)$ ,  $\nu^2 = \sigma^2/(1-\rho^2)$ , so  $\mu$  and  $\nu^2$  are the unconditional mean and variance of  $y_t$  according to the representation (3), then the Jacobian of the transformation from  $(c, \rho, \sigma^2)$  to  $(\mu, \rho, \nu^2)$  is

$$\left| \frac{\partial(\mu, \rho, \nu^2)}{\partial(c, \rho, \sigma^2)} \right| = \begin{vmatrix} \frac{1}{1-\rho} & 0 & 0 \\ \frac{c}{(1-\rho)^2} & 1 & \frac{2\sigma^2\rho}{(1-\rho^2)^2} \\ 0 & 0 & \frac{1}{1-\rho^2} \end{vmatrix} = \frac{1}{(1-\rho)(1-\rho^2)} . \quad (4)$$

Thus a flat prior on  $(\mu, \rho, \nu^2)$ , which does not seem a prior unreasonable, implies a prior proportional to the right-hand side of (4) on  $(c, \rho, \sigma^2)$ , i.e. a prior that puts heavy weight on values of  $\rho$  near 1. And of course use of OLS estimates and standard errors in raw form, which corresponds to

---

\*Copyright 2000 by Christopher A. Sims. This document may be reproduced for educational and research purposes, so long as the copies contain this notice and are retained for personal use or distributed free.

using a flat prior on  $c, \rho, \sigma^2$ , is equivalent to using a prior pdf on  $(\mu, \rho, \nu^2)$  that is close to zero in a neighborhood of  $\rho = 1$ .

If  $\rho = 1$ , then there is no representation of the form (3), but we can write

$$y_t = y_0 + ct + \sum_{s=0}^{t-1} \varepsilon_{t-s}. \quad (5)$$

Note that this means that when  $\rho = 1$ , the coefficient  $c$  switches from telling us about the level of the mean, to telling us about the slope of a linear trend line. A flat prior on  $c$  then means not just that we don't know the level of the  $y$  process, but that we believe that trends with arbitrarily large rates of increase or decrease are a priori not unlikely.

**Univariate AR with constant—Implausible initial conditions:** When we estimate this model by OLS, we are conditioning on  $y(0)$ , i.e. assuming that its distribution is unrelated to the parameters  $c, \rho, \sigma^2$ . There is a strong tendency for OLS estimates  $\hat{c}, \hat{\rho}, \hat{\sigma}^2$  to be such that

$$\frac{y(0) - \hat{c}/(1 - \hat{\rho})}{\hat{\sigma}/\sqrt{1 - \hat{\rho}^2}} > 2, \quad (6)$$

in other words such that the initial value of  $y$  is more than 2 standard errors away from its mean value. For example, Figure 1 (drawn from Sims (2000), on the reading list) shows results from estimating this model from 100 computer-generated time series on  $y$ , based on  $\rho = 1, c = 0$ . The 17 displayed plots are the 17 for which the left-hand-side of (6) is the largest. All these estimates show  $\hat{\rho} < 1$ , some of them much less than one. What is happening is that trend or initial curvature in the time path of the random walk data is being “explained” as due to a  $y_0$  far from its mean, so that the data reflect a steady exponential return to the mean.

In most applications, it is implausible that  $y_0$  is so many standard-deviations away from  $\mu$ . It makes sense, then, to use a prior that reflects our beliefs about this. But note that there is a connection between the classical bias of OLS — its tendency to produce estimates less than one in absolute value even when  $\rho = 1$  — and the tendency for estimates to imply big  $y_0 - \mu$  and big initial transients. Classical unit root econometrics, by focusing attention on the possibility that  $\rho = 1$  and suggesting the practice of testing  $H_0 : \rho = 1$ , accepting it unless there is strong evidence against it, might be interpreted as an indirect way to counteract the tendency of OLS to produce estimates with large initial transients.

**Multivariate VAR's can fit high-order polynomials:** In general, a VAR with  $k$  lags on each of  $m$  variables, with no constant terms, can fit exactly data consisting of a vector of  $m$  polynomials in  $t$ , each of order  $mk - 1$ .

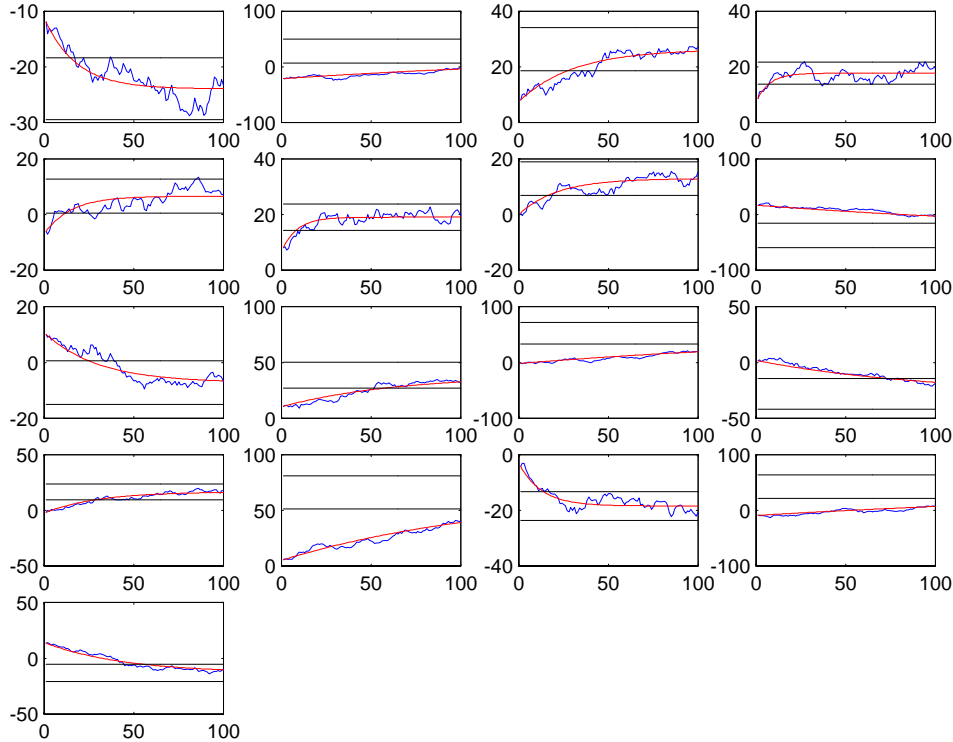


FIGURE 1. Initial Conditions Rogues' Gallery

Note: These are results from 17 of 100 draws of 100 terms from the process  $\{y_t | y_{t-1}\} \sim \text{i.i.d. } N(0, 1)$ . Blue lines are Monte Carlo data. Red lines are deterministic components, computed as  $(1 - \hat{\rho}^t)\hat{c}/(1 - \hat{\rho}) + \hat{\rho}^t y_0$ . Horizontal lines are 95% probability bands around the unconditional mean.

This should be plausible, because each polynomial has  $mk$  distinct coefficients, and there are  $m$  such polynomials, for a total of  $m^2k$  coefficients. This is exactly the number of free coefficients in an unrestricted VAR with lag length  $k$  and  $m$  variables. [Though if the coefficients of the polynomial are chosen “at random” there is probability zero that they cannot be fit perfectly, there are special-case counterexamples. For example, if  $y_1(t) = t^3$  and  $y_2(t) \equiv 1$ , then this being a pair of order-3 polynomials, we might think we could fit them with a bivariate VAR of order 2 (so  $mk - 1 = 3$ ). But we can't. A necessary condition on the polynomials is that the matrix of coefficients on the  $m$  highest-order terms in the polynomials must be non-singular.]

It is common in applied work to use 6 or more variables and 5 or more lags. A system like that can fit, without residual error, random collections of 29th-order polynomials. Of course few actual economic time series are exactly polynomials in  $t$ , but the fact that a VAR of standard form can

produce such complex deterministic behavior is something we must be aware of in considering estimation results. If a 6-variable, 5-lag system is estimated from 1950-2000 quarterly economic data by OLS, it is very likely that the estimates will imply that rather complicated patterns of behavior in the data were predictable from the 1950 initial conditions. Unless we believe such results are plausible, we should use priors that downweight this possibility. In multivariate models OLS estimates have an even stronger tendency than in univariate models to imply that initial conditions were unusual and deterministic trend behavior accounts for an implausibly large part of historical variation. Graphs in Sims (revised 1996) from the reading list illustrate this.

## 2. CLASSICAL APPROACHES TO A SOLUTION

OLS estimates in autoregressive models do not have normal distributions, even when equation disturbances are normal, as you should recall from earlier econometrics courses. In an AR1 like (1), OLS estimates of  $\rho$ , though not exactly normal, are asymptotically normal if  $|\rho| < 1$ , as you should also have seen in an earlier course. This justifies the usual estimates and test statistics as asymptotic approximations when  $|\rho| < 1$ . When  $\rho = 1$ , there is still a limiting distribution, but it is not normal. This discontinuous change at  $\rho = 1$  in the nature of the limiting distribution makes this version of the classical theory difficult to use in applied work. If one attempted to use it literally to form a confidence interval, for example, it would lead to disconnected intervals, most commonly to a confidence set consisting of an interval about  $\hat{\rho}$  that does not include one, together with the point  $\rho = 1$ . (This point is discussed in the “Helicopter Tour” paper on the reading list (Sims and Uhlig, 1991).)

There is a classical approach to avoiding these artificial discontinuities. One carries out a “local to unity” asymptotic analysis, in which it is assumed that  $\rho = 1 - c/T$ . This kind of asymptotic approximation is not discontinuous at  $\rho = 1$ . But the existence of two competing types of asymptotics, with no available guidance as to which type to use, is itself a problem for applied work. One criterion for using local-to-unity asymptotics might be a prior judgment that this is likely to be a good approximation, in which case why not be explicit about invoking prior information and take a Bayesian approach? Another criterion might be to choose local-to-unity asymptotics because sample information tells us that  $\rho$  is close to one. But if this is how we proceed, our confidence intervals arise as a complicated two-step function of the data; the usual theory, which ignores the randomness in what type of asymptotics is used to generate the distribution theory, gives an inaccurate picture of the sampling properties of the statistics.

### 3. A BAYESIAN APPROACH TO A SOLUTION

**3.1. Dummy observation priors.** The definition of a conjugate prior is that it is a pdf for the parameters that has the same shape as the likelihood function from a sample generated by the model. This fact can be helpful in computation, as it means that we can always formulate the prior as a set of “dummy observations” that are added at the end of the sample. This allows us to trick a standard statistical package that applies OLS into providing us with a Bayesian posterior. But the dummy observation idea is also helpful to us in formulating priors. Particularly in complicated multivariate AR models, it is likely to be easier to formulate dummy observations, which in a sense describe what we believe about likely behavior of the data, than it is to formulate priors on the AR parameters themselves.

In a Normal linear regression model,  $\{y | \beta\} \sim N(X\beta, \sigma^2 I)$ , as we have already seen, the likelihood has the shape of a normal-inverse-gamma distribution. That is, treating the likelihood as a joint pdf for  $\sigma^2, \beta$  implies that the marginal pdf for  $\frac{1}{2}\hat{u}'\hat{u}\sigma^{-2}$  is  $\Gamma(\frac{1}{2}t - 1)$  and the conditional pdf for  $\{\beta | \sigma\}$  is  $N(\hat{\beta}, \sigma^2(X'X)^{-1})$ , where  $\hat{u}$  is the vector of OLS residuals and  $\hat{\beta}$  is the OLS estimate of  $\beta$ . So given any prior pdf for  $\beta$  that can be expressed as  $\{\beta | \sigma\} \sim N(\mu, \sigma^2\Omega)$ , which is the conjugate-prior form, we can express it equivalently as a set of dummy observations  $y^*$  on  $y$  and  $X^*$  on  $X$  of the form

$$y^* = W\mu, \quad X^* = W^{-1}, \quad (7)$$

where  $W$  is some nonsingular matrix such that  $WW' = \Omega$ .

It may seem unreasonable to specify a prior on  $\beta$  whose standard deviation is scaled by the residual standard deviation of the model, and in fact it often is. This is a disadvantage of the conjugate prior. It can be dealt with in practice numerically, for example by integrating numerically over  $\sigma$ . But in time series applications, as we see below, the scaling of the prior by the residual variance may not be so unreasonable.

**3.1.1. The univariate AR case.** A general treatment of priors in VAR's is in Sims and Zha (1998). Here we consider the simplest possible case to develop intuition. We have the model (2), and we want to formulate a convenient prior that embodies our disbelief in models like those shown in Figure 1. This can be done with a pair of dummy observations, the first of which embodies the idea that a prediction of  $y_{t+1} = y_t$  is likely to be accurate, in the sense of having residual variance  $\sigma^2$ . This can be accomplished in a straightforward way: set  $X^* = [1 \ y^*]$ . There is of course a question of what level to set  $y^*$  at, but since we are conditioning on  $y_0$ , a natural candidate is  $y^* = y_0$ . We might at the same time want to put a very diffuse prior on  $c$ , for example letting its standard deviation be 100 times that of  $\varepsilon_t$ . This can be done with a dummy observation in which  $y^* = 0, X^* = [.01 \ 0]$ .

These two dummy observations imply the prior

$$\begin{bmatrix} \rho \\ \sigma^2 \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \sigma^2 10^4 \begin{bmatrix} 1 & -1/y_0 \\ 1.0001/y_0^2 & \end{bmatrix} \right). \quad (8)$$

The mean vector here shows that the implied prior centers on a pure unit root model, with no constant term, but the variances of both  $\rho$  and  $c$  are high. What the prior determines with some precision (from the first dummy observation) is  $c + \rho y_0$ . As  $\rho \rightarrow 0$ ,  $c$  is implied to get close to  $y_0$ . As  $\rho \rightarrow 1$ ,  $c$  is implied to get close to zero. So the prior expresses skepticism about models in which  $y_1$  is projected deterministically, based on  $y_0$ , to be very different from  $y_0$ , which will tend to push estimates away from ones like those in Figure 1.

Note that the second of the two dummy observations here is likely to have almost no effect on results. We could just omit it altogether, so we have only one dummy observation. Then obviously the formulas (7) do not apply directly. However, using a less than full rank set of dummy observations is a reasonable practice, corresponding to using a flat prior on some dimensions of the parameter space.

Note also that we could use more than two dummy observations. For example, we might like to express a belief that  $\rho$  is not too far from zero. We can do so by just adding an additional dummy observation. A prior with three dummy observations on two parameters can always be expressed equivalently in terms of two dummy observations, but often (as in this example) it can be easier to understand the prior expressed in terms of dummy observations than it is to understand the implications of the mean and variance of the implied joint normal distribution.

#### 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, [trends/asapaper.pdf](#).
- SIMS, C. A., AND H. D. UHLIG (1991): “Understanding Unit Rooters: A Helicopter Tour,” *Econometrica*, 59.
- SIMS, C. A., AND T. ZHA (1998): “Bayesian Methods for Dynamic Multivariate Models,” *International Economic Review*.