

Topics in Macroeconomics
Economics, 411
Christiano, Fall 2002.

The purpose of these notes is to discuss the Bayesian perspective on the computation of confidence intervals for vector autoregressions. The notes summarize the discussion in Sims and Uhlig (1991).¹ That paper is in fact very accessible and the student is encouraged to study it. Looking at these notes first might be helpful because I have not skipped any steps in the argument.

The argument is that Bayesian and classical methods for characterizing parameter uncertainty can lead to quite different results in a time series context, and that surely any reasonable person prefers the Bayesian method. (This is Sims-Uhlig speaking, not necessarily me!)

My discussion only treats the case of a scalar first order autoregression. This discussion is itself divided into two parts. The first treats the innovation variance as a known constant. The second treats it as an unknown. The Bayesian posterior distribution for the case of a general vector autoregression with finite lags is nicely derived in Evans and Marshall (2002) (see also Doan (2000) for a statement of the posterior, and Zellner (1971, pp. 224-227) for a derivation in a closely related context.)

We consider the following data generating process:

$$y_t = \mu + \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma), \quad \mu = 0, \quad y_0 = 0, \quad \sigma = 1. \quad (0.1)$$

The econometrician observes a sample of $T = 100$ observations from this process and wishes to infer the value of ρ . To begin, I will assume that the econometrician knows the true value of μ to be zero, and also knows the true value of the variance. Throughout, the econometrician knows that ε_t is normally distributed and that the initial condition, $y(0)$, is a non-random variable.

1. Mean, Variance Known

The likelihood function of y_1, \dots, y_T , conditional on ρ , can be built up as the product of the sequence of conditional likelihoods. Thus,

$$L(y_1; \rho) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \frac{(y_1 - \rho \times 0)^2}{\sigma^2} \right\},$$

¹It also summarizes a part of the discussion in Sims and Zha (1998). However, no attempt is made to review the main idea in that paper, which is to explain how to compute confidence intervals in VAR's when there are overidentifying restrictions.

Also,

$$\begin{aligned}
L(y_2, y_1; \rho) &= L(y_2|y_1; \rho)L(y_1; \rho) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 \exp\left\{-\frac{1}{2}\frac{(y_2 - \rho y_1)^2}{\sigma^2}\right\} \times \exp\left\{-\frac{1}{2}\frac{(y_1 - \rho \times 0)^2}{\sigma^2}\right\} \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 \exp\left\{-\frac{1}{2}\sum_{t=1}^2 \frac{(y_t - \rho y_{t-1})^2}{\sigma^2}\right\}.
\end{aligned}$$

In this way, the likelihood of a sample, y_1, \dots, y_T is:

$$L(y_T, \dots, y_2, y_1; \rho) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^T \exp\left\{-\frac{1}{2}\sum_{t=1}^T \frac{(y_t - \rho y_{t-1})^2}{\sigma^2}\right\},$$

with the understanding, $y_0 = 0$.

We now view the parameter value, ρ , as unknown, with prior distribution, $p(\rho)$. Then, the conditional distribution of ρ , after seeing $y = y_1, \dots, y_T$ is:

$$p(\rho|y) = \frac{p(\rho, y)}{p(y)} = \frac{L(y; \rho)p(\rho)}{p(y)}.$$

If we treat the data as given, then $p(y)$ is just a number. In addition, we suppose that the prior over ρ is diffuse, so that $p(\rho) = 1$. In this case,

$$p(\rho|y) \propto L(y; \rho),$$

where, as usual, ‘ \propto ’ signifies ‘is proportional to’, where the constant factor of proportionality can be found by scaling so that the integral over ρ equals unity.

It is convenient to take into account a particular identity, when writing out $p(\rho|y)$:

$$\begin{aligned}
\sum_{t=1}^T (y_t - \rho y_{t-1})^2 &= \sum_{t=1}^T (y_t - \hat{\rho} y_{t-1} + \hat{\rho} y_{t-1} - \rho y_{t-1})^2 \\
&= \sum_{t=1}^T \left((y_t - \hat{\rho} y_{t-1})^2 + (\hat{\rho} - \rho)^2 y_{t-1}^2 + 2(y_t - \hat{\rho} y_{t-1}) y_{t-1} (\hat{\rho} - \rho) \right) \\
&= \sum_{t=1}^T (y_t - \hat{\rho} y_{t-1})^2 + (\hat{\rho} - \rho)^2 \sum_{t=1}^T y_{t-1}^2 + 2(\hat{\rho} - \rho) \sum_{t=1}^T (y_t - \hat{\rho} y_{t-1}) y_{t-1}.
\end{aligned}$$

Now, suppose that $\hat{\rho}$ is the ordinary least squares estimator of ρ , imposing $\mu = 0$. Then,

$$\sum_{t=1}^T (y_t - \hat{\rho}y_{t-1}) y_{t-1} = 0,$$

so that

$$\hat{\rho} = \frac{\sum_{t=1}^T y_t y_{t-1}}{\sum_{t=1}^T y_{t-1}^2}$$

In this case, the posterior distribution can be written:

$$\begin{aligned} p(\rho|y) &\propto \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^T \exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{t=1}^T (y_t - \hat{\rho}y_{t-1})^2 + (\hat{\rho} - \rho)^2 \sum_{t=1}^T y_{t-1}^2 \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \frac{(\hat{\rho} - \rho)^2}{s^2} \right\}, \end{aligned}$$

where

$$s^2 = \frac{1}{\sum_{t=1}^T y_{t-1}^2}.$$

In the expression for $p(\rho|y)$, terms involving only y have been dropped, and σ is set to unity. Thus, the posterior distribution of ρ is normal, with variance s^2 and mean $\hat{\rho}$.

Interestingly, this is the distribution for $\hat{\rho}$ one obtains from classical asymptotic analysis. There it is $\hat{\rho}$ that is treated as the random variable, and the notion is that it has a normal distribution with mean ρ (the ‘true value’) and sample standard deviation, s . It is typical to construct confidence intervals, by supposing that $(\hat{\rho} - \rho)/s$ has a normal distribution. This leads to the 95% ‘confidence interval’,

$$\hat{\rho} - 1.96 \times s \leq \rho \leq \hat{\rho} + 1.96 \times s,$$

which, in repeated samples, should contain the true value 95% of the time. The Bayesian does not think in terms of repeated samples, only in terms of the sample actually available. From the perspective of the Bayesian, it is the true value of ρ that is uncertain.² In the Bayesian interpretation, the above (‘Bayesian posterior probability’) interval is a way to characterize the posterior distribution, when the prior is uniform. The Bayesian says that the true parameter lies inside this interval

²Presumably, the Bayesian believes that there is a unique, true value of ρ . The randomness about ρ in the posterior distribution characterizes the Bayesian’s degree of belief about the different possible values of ρ .

with 95% probability. In many cases, the distinction between what the Bayesian says about this interval and what the classical econometrician says is very slight. The two part ways when the classical econometrician believes that the asymptotic theory for stationary processes is a poor approximation. This happens when, for example, unit roots enter the picture. The Bayesian attributes no special significance to these, while the classical econometrician invokes a discontinuously different asymptotic sampling theory for this case. The Bayesian continues to compute the interval in the same way (as long as the normality of the errors is still maintained!).

A clear illustration of the difference between the Bayesian ‘confidence interval’ and the classical one, is described in the example in Sims and Uhlig (1991). They show how it is that when $\hat{\rho}$ is observed, a classical p value for the null hypothesis that true $\rho = 1$ will be larger than one for $\rho = 0.90$. For this example, they report that the p value for $\rho = 1$ is 0.12 while it is 0.04 for $\rho = 0.90$. Thus, a classical econometrician would reject, at the 5% level, the hypothesis of $\rho = 0.9$ and easily accept the other. At the same time, the Bayesian will assign equal probability to each. Sims and Uhlig argue that any ‘reasonable’ person will take the position of the Bayesian. So, what is going wrong, then, with the classical p values?³

Sims and Uhlig argue that the classical p values are being distorted by ‘irrelevant information’. They grant that if $\rho = 1$, then the likelihood of observing $\hat{\rho}$ much below 0.95 is greater than the likelihood of observing $\hat{\rho}$ much above 0.95 when $\rho = 0.9$. They say, however, that ‘...for deciding what [a] sample tells us about ρ , the implications of the competing hypotheses about $\hat{\rho}$ ’s we have not observed are irrelevant.’

Figure 1 is useful for visualizing the Sims-Uhlig point. I suppose that the econometrician has observed 0.95 in a data set generated by the data generating mechanism described above. The econometrician only knows $\mu = 0$ and the distribution of ε_t . Following Sims-Uhlig, I also suppose that the only thing the econometrician observes about the data is $\hat{\rho}$. In particular, the data set itself, y_1, \dots, y_{100} , is *not* observed. This assumption has the important practical advantage of reducing the dimension of the problem sufficiently so that results can be exhibited graphically. Numerical methods have to be used to compute the posterior distribution since the assumption that y_1, \dots, y_{100} is not observed means that

³Recall how the p - values are constructed. For the null hypothesis, $\rho = 1$, the p - value is the probability, under $\rho = 1$, of obtaining a $\hat{\rho}$ even smaller than 0.95, which is the value presumably obtained in the sample. For $\rho = .8$, the p -value is the probability of obtaining $\hat{\rho}$ even bigger than 0.95.

the formula for the posterior distribution derived above does not apply. To compute the posterior distribution, I used the numerical method applied in Sims-Uhlig. The posterior distribution is displayed in Figure 1. It indicates the probability that $\hat{\rho} = 0.95$ (the presumed observed value) conditional on the various values of ρ on the horizontal axis being true. Note the (approximate) symmetry of the distribution. In particular, the area under the posterior distribution to the right of unity and to the left of 0.90 are roughly the same.⁴

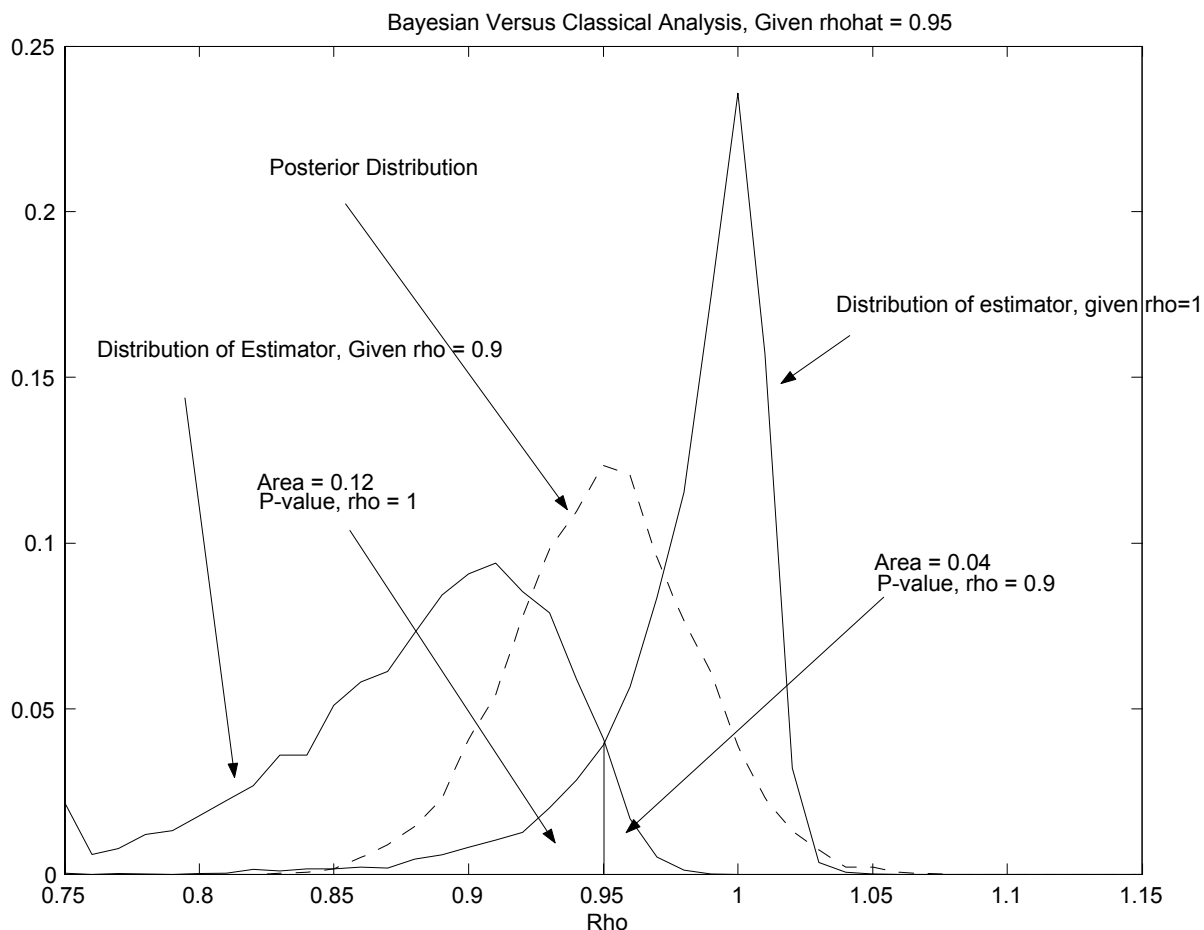
Also graphed in Figure 1 are the distributions of $\hat{\rho}$ conditional on $\rho = 1$ and $\rho = 0.9$.⁵ Note that the Bayesian assigns equal posterior probability to $\rho = 0.9$ and $\rho = 1.0$. At the same time, the p -value associated with the hypothesis, $\rho = 1$, is far greater than the p -value associated with the hypothesis, $\rho = 0.9$. Standard application of classical p -values would assign a higher likelihood to $\rho = 1$ than to $\rho = 0.9$. Clearly, that would be a mistake in this case. It is hard not to agree with Sims-Uhlig that the reasonable person would follow the Bayesian procedure

⁴The probability of any given value of ρ , given $\hat{\rho} = 0.95$, is computed as follows. For each ρ , I generated 10,000 data sets using that value of ρ , from the data generating mechanism specified above. In each data set I computed $\hat{\rho}$. I then computed the frequency of times that a simulated value of $\hat{\rho}$ fell in the interval $[0.945, 0.955]$. The values of ρ considered are 0.75, 0.76, ..., 1.15.

⁵I computed these distribution by generating 10,000 artificial sets of length 100 each from the above time series model, with the indicated value of ρ . I then computed the histogram of the resulting values of $\hat{\rho}$. These histograms are reported in the figure. In both cases, the probabilities are normalized to sum to unity.

and assign equal plausibility to the notion that $\rho = 0.9$ and $\rho = 1.0$.

Figure 1



Since the distribution of $\hat{\rho}$ given $\rho = 0.9$ and $\rho = 1.0$ are so different, it is worth discussing this a bit further. In particular, note how the distribution of $\hat{\rho}$ is both more concentrated (i.e., has less sampling uncertainty) and more biased when $\rho = 1$, than when $\rho = 0.9$. Fundamentally, this is why the posterior likelihood of the the two values of ρ are the same. On the one hand, bias considerations alone make $\rho = 1$ more likely than $\rho = 0.9$ when $\hat{\rho} = 0.95$ is observed. On the other hand, sampling considerations alone make 0.9 more likely. As it happens, the two types of considerations cancel exactly in this example.

It is of interest to understand why the bias of $\hat{\rho}$ is downward, and why this is an increasing function of ρ . This fact is well known, at least since Hurwicz (1950).

To see the bias, suppose $T = 3$, so that

$$\begin{aligned}\hat{\rho} &= \frac{y_3y_2 + y_2y_1}{y_2^2 + y_1^2} = \frac{(\rho y_2 + \varepsilon_3)y_2 + (\rho y_1 + \varepsilon_2)y_1}{y_2^2 + y_1^2} \\ &= \rho + \frac{\varepsilon_3y_2 + \varepsilon_2y_1}{y_2^2 + y_1^2} \\ &= \rho + \left(\frac{y_2}{y_2^2 + y_1^2}\right)\varepsilon_3 + \left(\frac{y_1}{y_2^2 + y_1^2}\right)\varepsilon_2.\end{aligned}$$

In the standard regression context, the objects in parentheses are independent of the the error terms. This is the basis for the standard result that least squares is unbiased, $E\hat{\rho} = \rho$. However, here the second expression in parentheses is not independent of the error term that it multiplies. The problem lies with the denominator term, which contains y_2 , which in turn is a function of ε_2 . The numerator term is obviously not a source of the problem. As a result,

$$E\left(\frac{y_1}{y_2^2 + y_1^2}\right)\varepsilon_2 \neq E\left(\frac{y_1}{y_2^2 + y_1^2}\right)E\varepsilon_2.$$

By considering the case of general T , it is easy to see why this problem gets smaller as $T \rightarrow \infty$. In the general case, the expression of interest is:

$$E\left(\frac{y_{t-1}}{\sum_{j=1}^T y_{j-1}^2}\right)\varepsilon_t.$$

As before, the ‘problem’ lies in the denominator, where the sum of squares of all the data appear. Under the assumption of covariance stationarity, ε_t has an impact (is correlated with) only a limited number of future y_t ’s. As $T \rightarrow \infty$, these play a vanishing role in the sum. Since the numerator in the ratio is not a problem in any case, it follows that as $T \rightarrow \infty$

$$E\left(\frac{y_{t-1}}{\sum_{j=1}^T y_{j-1}^2}\right)\varepsilon_t \rightarrow E\left(\frac{y_{t-1}}{\sum_{j=1}^T y_{j-1}^2}\right)E\varepsilon_t = 0.$$

The problem just described is obviously greater for larger values of ρ . When ρ is large, ε_t is correlated with a greater number of elements in the sum, $\sum_{j=1}^T y_{j-1}^2$, and so it takes a greater value of T before this correlation becomes unimportant.

This is a heuristic discussion of the Hurwicz bias, and the reason why it is smaller for smaller ρ . It explains why the distribution of $\hat{\rho}$ exhibits a greater left-shift in Figure 1 when $\rho = 1$ than when $\rho = 0.9$.

2. Only the Variance is Known

We now suppose that the value of μ is not known. In this case, the likelihood function is:

$$L(y_T, \dots, y_2, y_1; \mu, \rho) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^T \exp \left\{ -\frac{1}{2} \sum_{t=1}^T \frac{(y_t - \mu - \rho y_{t-1})^2}{\sigma^2} \right\}$$

The OLS estimator of the parameters is defined by:

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{\mu} - \hat{\rho} y_{t-1}) &= 0 \\ \sum_{t=1}^T (y_t - \hat{\mu} - \hat{\rho} y_{t-1}) y_{t-1} &= 0 \end{aligned} \quad (2.1)$$

Then,

$$\begin{aligned} & \sum_{t=1}^T (y_t - \mu - \rho y_{t-1})^2 \\ &= \sum_{t=1}^T (y_t - \hat{\mu} - \hat{\rho} y_{t-1} + \hat{\mu} + \hat{\rho} y_{t-1} - \mu - \rho y_{t-1})^2 \\ &= \sum_{t=1}^T [(y_t - \hat{\mu} - \hat{\rho} y_{t-1})^2 + (\hat{\mu} + \hat{\rho} y_{t-1} - \mu - \rho y_{t-1})^2 \\ & \quad + 2(y_t - \hat{\mu} - \hat{\rho} y_{t-1})(\hat{\mu} - \mu) + 2(y_t - \hat{\mu} - \hat{\rho} y_{t-1})(\hat{\rho} - \rho)y_{t-1}] \\ &= \sum_{t=1}^T [(y_t - \hat{\mu} - \hat{\rho} y_{t-1})^2 + (\hat{\mu} + \hat{\rho} y_{t-1} - \mu - \rho y_{t-1})^2] \\ &= \sum_{t=1}^T (y_t - \hat{\mu} - \hat{\rho} y_{t-1})^2 + \sum_{t=1}^T ((\hat{\mu} - \mu) + (\hat{\rho} - \rho)y_{t-1})^2 \\ &= \sum_{t=1}^T (y_t - \hat{\mu} - \hat{\rho} y_{t-1})^2 + T(\hat{\mu} - \mu)^2 + (\hat{\rho} - \rho)^2 \sum_{t=1}^T y_{t-1}^2 + 2(\hat{\mu} - \mu)(\hat{\rho} - \rho) \sum_{t=1}^T y_{t-1} \end{aligned} \quad (2.2)$$

The last three expressions can be written:

$$\begin{aligned} & \begin{pmatrix} (\hat{\mu} - \mu) & (\hat{\rho} - \rho) \end{pmatrix} \begin{bmatrix} T & \sum_{t=1}^T y_{t-1} \\ \sum_{t=1}^T y_{t-1} & \sum_{t=1}^T y_{t-1}^2 \end{bmatrix} \begin{pmatrix} (\hat{\mu} - \mu) \\ (\hat{\rho} - \rho) \end{pmatrix} \\ &= \begin{pmatrix} (\hat{\mu} - \mu) & (\hat{\rho} - \rho) \end{pmatrix} X'X \begin{pmatrix} (\hat{\mu} - \mu) \\ (\hat{\rho} - \rho) \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} X &= [\tau, \tilde{y}], \\ \tau &= \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y_{T-1} \\ \vdots \\ y_0 \end{pmatrix} \end{aligned} \tag{2.3}$$

So, the likelihood can be written

$$L(y; \mu, \rho) \propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} \hat{\mu} - \mu & \hat{\rho} - \rho \end{pmatrix} X'X \begin{pmatrix} \hat{\mu} - \mu \\ \hat{\rho} - \rho \end{pmatrix} \right\}.$$

Again, this looks like the classical asymptotic distribution, with ρ, μ playing the role of the ‘true’ parameter values and

$$\begin{pmatrix} \hat{\mu} - \mu & \hat{\rho} - \rho \end{pmatrix} \sim N(0, (X'X)^{-1}).$$

3. Mean, Variance Unknown

Now, we include all three parameters in the likelihood function, μ, ρ and σ :

$$L(y_T, \dots, y_2, y_1; \rho, \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^T \exp \left\{ -\frac{1}{2} \sum_{t=1}^T \frac{(y_t - \mu - \rho y_{t-1})^2}{\sigma^2} \right\}.$$

Substituting from (2.2),

$$\begin{aligned} &L(y_T, \dots, y_2, y_1; \rho, \mu, \sigma) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^T \exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{t=1}^T (y_t - \hat{\mu} - \hat{\rho} y_{t-1})^2 + \begin{pmatrix} \hat{\mu} - \mu & \hat{\rho} - \rho \end{pmatrix} X'X \begin{pmatrix} \hat{\mu} - \mu \\ \hat{\rho} - \rho \end{pmatrix} \right] \right\}, \end{aligned}$$

where X is defined in (2.3).

4. Exercise

When the constant term in (0.1) is treated as an unknown to be estimated, then the OLS estimator of $\rho, \hat{\rho}$, is given by (2.1). The impact on the estimator of $\hat{\rho}$ may be seen by first solving the first equation for $\hat{\mu}$:

$$\hat{\mu} = \bar{y} - \hat{\rho}\bar{y}_{-1},$$

where

$$\bar{y} = \frac{1}{T} \sum_{t=1}^T y_t, \quad \bar{y}_{-1} = \frac{1}{T} \sum_{t=1}^T y_{t-1}.$$

Substituting out for $\hat{\mu}$ into the second equation in (2.1), and solving the result for $\hat{\rho}$, we obtain:

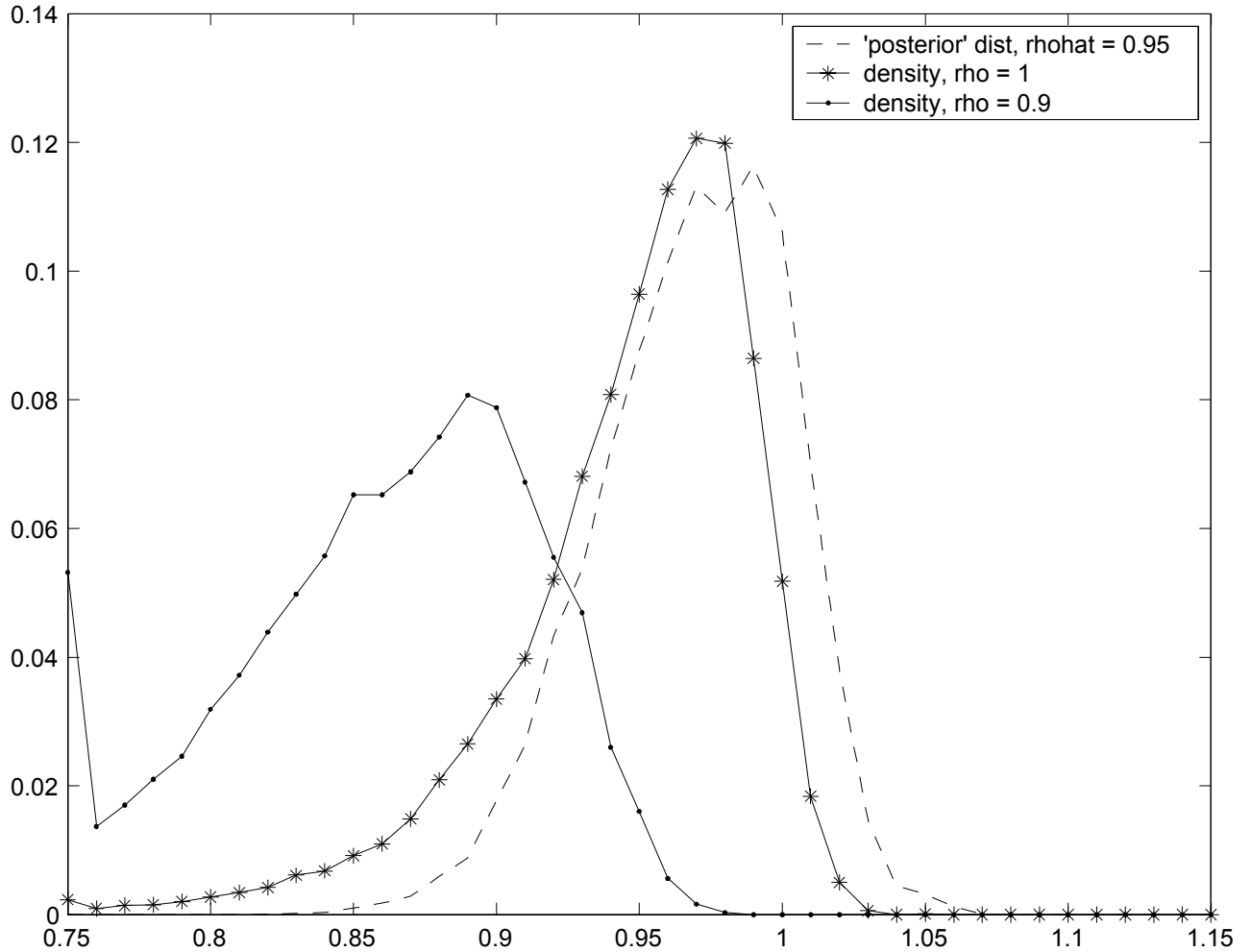
$$\hat{\rho} = \frac{\sum_{t=1}^T (y_t - \bar{y})(y_{t-1} - \bar{y}_{-1})}{\sum_{t=1}^T (y_{t-1} - \bar{y}_{-1})^2}$$

Figure 2 redoes the experiments reported in Figure 1. In computing the ‘posterior’ probability of various values of ρ , I imposed $\mu = 0$ but used the formula for $\hat{\rho}$ given above.

There are several things worth noting about the Figure 2. First, notice how much more pronounced the bias in $\hat{\rho}$ is when $\rho = 0.9$ or $\rho = 1$. Second, notice how

the ‘posterior distribution’ is now shifted to the left.

Figure 2



1. Explain why the ‘posterior distribution’ that appears in Figure 2 is inconsistent with the discussion in the preceding part of the text. Explain why it is fact misleading to think of the above ‘posterior distribution’ as an actual posterior distribution?
2. Explain how removing the sample mean from the data before computing $\hat{\rho}$ adds a second channel for sample bias, a channel that is very similar to the Hurwicz bias discussed above. Explain how that source of bias vanishes as $T \rightarrow \infty$.

References

- [1] Doan, Thomas, 2000, *RATS User's Guide*, Version 5, Evanston, IL.
- [2] Evans, Charles and David Marshall, 2002, 'Identifying Structural Vector Autoregressions Using Noisy Shock Measures: Bayesian Inference', manuscript, Federal Reserve Bank of Chicago.
- [3] Hurwicz, Leo, 1950, 'Least Squares Bias in Time Series,' in *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph no. 10, ed. by T. C. Koopmans, New York: Wiley.
- [4] Sims, Christopher and Harald Uhlig, 1991, 'Understanding Unit Rooters: A Helicopter Tour,' *Econometrica*, vol. 59, issue 6.
- [5] Sims, Christopher and Tao Zha, 1998, 'Error Bands for Impulse Responses', manuscript on Sims' web page (also published in *Econometrica*).
- [6] Zellner, A., 1971, *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley & Sons, Inc.