PRIORS FOR THE AR(1) MODEL

Parameterization Issues and Time Series Considerations

PETER C. SCHOTMAN
University of Limburg

Two issues have come up in the specification of a prior in the Bayesian analysis of time series with possible unit roots. The first issue deals with the singularity that is due to the local identification problem of the unconditional mean of an AR(1) process in the limit of a random walk. This singularity problem is related to the difference between a structural parameterization and the linear reduced form in a standard regression model with fixed regressors. The second is related to the time series nature of the regressor in an AR(1) model. In this paper we will concentrate on the parameterization issue. First, it is shown that the posterior of the autoregressive parameter can be very sensitive to the degree of prior dependence between the unconditional mean and the autocorrelation parameter. Second, the time series nature of the problem suggests a particular form of this dependence.

1. INTRODUCTION

Two issues separate the Bayesian analysis of autoregressive time series models from standard Bayesian inference in linear regression models. The first difference is the time series nature of the problem. In an influential paper on Bayesian methods in time series models, Phillips [2] argued a case for "objective ignorance" priors, which, as he showed, can be very different from the usually employed flat priors. The Jeffreys prior that Phillips [2] derived for the AR(1) model with intercept and trend explicitly accounts for the time series properties of an AR(1) process, whether it is stationary or not.

The second important difference is the inherent nonlinearity in the parameterization of the model. Under the unit root hypothesis, the unconditional mean of the AR(1) process does not exist, and this causes the constant term in the model to be (locally) unidentified. Schotman and van Dijk [4] and Zivot [14] considered a prior for the AR(1) model that was specifically designed to overcome the pathology in the posterior due to this singularity.

The first draft of the paper circulated under the title "Objective Information and Ignorance," and was written while I visited the Woodrow Wilson School of Public and International Affairs at Princeton University. I thank Harald Uhlig and Herman van Dijk for some helpful discussions, and three anonymous referees and the editor for their comments on the first draft of this paper. All errors are, of course, my own.
As indicated by Phillips [3], the problems of Bayesian inference in dealing with this singularity are not specific to the unit root hypothesis but will arise in many other structural models. In fact, the singularity problem is analogous to determining the ratio of two regression coefficients, a problem studied by Zellner [13].

In the first part of this paper we will investigate the sensitivity of the posterior with respect to alternative types of priors in a simple fixed regressor model with this singularity. We introduce a class of priors that is motivated both by the interpretation of parameters of interest and by the priors that have been suggested for the AR(1) model. The analysis of the singularity prepares for the second purpose of the paper. Because the various priors that have been proposed for the AR(1) model are all within this class of priors, we can analytically compare the different behaviors of the resulting posteriors. In particular, it will be possible to explain why the "critics" prior of Phillips [2] and the conditional normal prior of Schotman and van Dijk [4] often give similar posteriors in the stationary region, and how alternative assumptions regarding the initial condition affect the posterior.

The paper is organized in two main sections. Section 2 deals with the singularity of the parameterization. Section 3 deals with the peculiarities of the AR(1) model with unknown mean. The paper finishes with some concluding remarks.

2. A REGRESSION MODEL WITH A SINGULARITY IN THE PARAMETERIZATION

The AR(1) model with unknown mean \( \mu \) can be written in several ways. Two representations have been contrasted in the recent discussion on the specification of a prior in the Bayesian analysis of the AR(1) model (see Zivot [14]). The reduced form of the model can be represented either in a form that is linear in the parameters,

\[
\Delta y_t = \alpha y_{t-1} + \delta + \epsilon_t, \tag{1}
\]

or, alternatively, in a form that is nonlinear in the parameters,

\[
\Delta y_t = \alpha (y_{t-1} - \mu) + \epsilon_t, \tag{2}
\]

where \( \alpha = -(1 - \rho) \), with \( \rho \) as the first-order autocorrelation parameter. The two models are different when \( \alpha = 0 \). In model (2) the intercept vanishes as \( \alpha \to 0 \), whereas the intercept is unrestricted in model (1). With the AR(1) interpretation in mind, model (2) is preferred because its parameters are interpretable and it preserves the structural restriction. Yet model (1) looks more tractable due to its linearity, and it is the form most often used.

Another important model with this singularity is the error correction mech-
anism (ECM), which, with only first-order dynamics, can be written in either of the two forms:

\[ \Delta y_t = \alpha y_{t-1} + \delta x_{t-1} + \epsilon_t \]  
\[ \Delta y_t = \alpha (y_{t-1} - \beta x_{t-1}) + \epsilon_t. \]

The parameter \( \beta \) determines the long-run relation between \( y \) and \( x \), and it is the parameter of interest in many economic applications. It is the parameter for which we will most likely be able to formulate an informative prior.

In this section we will concentrate on the singularity and not on time series aspects. We consider a regression model with fixed regressors,

\[ y = \alpha (x + \beta \iota) + e \]

where \( y, x, \iota, \) and \( e \) are vectors of length \( T \) containing the observations on the dependent variable, an independent variable, a constant term, and the error term respectively.\(^1\) The corresponding linear representation of the model is

\[ y = \alpha x + \delta \iota + e. \]

A flat prior \( p(\alpha, \delta, \sigma) \propto \sigma^{-1} \) on the parameters \( (\alpha, \delta, \ln \sigma) \) in (6) is equivalent to the prior \( p(\alpha, \beta, \sigma) \propto |\alpha| \sigma^{-1} \) for the parameters in (5). If \( \alpha \) is the only parameter of interest in this fixed regressor model, and if there is no subjective prior information on \( \alpha \), then there is little point in considering a more complicated prior than the flat prior on \( (\alpha, \delta) \). In the following discussion, we will therefore focus our attention on \( \beta \).

Inference on \( \beta \) will not be straightforward, however. Because \( \beta = \delta / \alpha \), we are dealing with the ratio of two regression coefficients. In a classical analysis the indirect least-squares estimator \( \hat{\beta} = \hat{\delta} / \hat{\alpha} \) will not have any integer moments in finite samples. The same is true in a Bayesian framework. With a flat prior on \( (\alpha, \delta) \), Zellner [13] showed that a special loss function is required for this problem to get sensible point estimates for \( \beta \). The problem is due to the behavior of the conditional posterior of \( \beta \) given \( \alpha \) close to \( \alpha = 0 \). To obtain existence of the posterior mean of \( \beta \), we can either sufficiently downweight the prior on \( \alpha \) close to \( \alpha = 0 \) or specify a proper prior on \( \beta \) for which the prior mean already exists. Given our interpretation of \( \beta \), we will take the latter route here. As we shall see, such a prior can dramatically affect the marginal posterior of \( \alpha \).

When \( \beta \) is a parameter of interest with some economic meaning, it is at least counterintuitive to put independent priors on \( \alpha \) and some hybrid reduced-form parameter \( \delta \). One alternative prior is a flat prior on the structural parameters \( (\alpha, \beta) \). Such a prior would simply specify our ignorance on \( \beta \). Zellner [12] argued that a flat prior could be regarded as being uninformative for any transformation of location parameters, such as \( \alpha, \beta, \) and \( \delta \), by adhering to the principle of Maximum Data Information Priors (MDIP's).
The MDIP is uninformative in the sense that it maximizes the information measure

\[ G = \int_y \int_\theta \ln \left\{ \frac{p(y|\theta)}{p(\theta)} \right\} p(y,\theta) \, d\theta \, dy \]  

(see Zellner [12]). This class of priors is not invariant with respect to nonlinear transformations of the parameters, as in (5) and (6). Formally, the MDIP is required to be a proper density. In this paper we will work with a class of priors that encompasses both the flat prior on \((\alpha, \beta)\) and the flat prior on \((\alpha, \delta)\) as special cases, and which also contains both proper and improper priors. The crucial part of the prior is the following conditional prior for \(\beta\) given \(\alpha\) and \(\sigma\):

\[ \beta | \alpha, \sigma \sim N \left( 0, \frac{\sigma^2}{K} \left| \alpha \right|^p \right), \]  

(7)

with \(K > 0\). Note that \(p = 0\) implies prior independence between \(\alpha\) and \(\beta\). For \(p = 2\) we have prior independence between \(\alpha\) and \(\delta\). The limit if \(K \to 0\) leads to the flat prior in either of the models (5) and (6). Both priors can be viewed as extreme cases. In the analysis of the AR(1) model in Section 3, we will see that various forms of the Jeffreys prior as well as the prior used by Schotman and van Dijk [4,6] all lead to \(p = 1\). The prior mean has been set to zero to simplify the algebra. The normal distribution and the dependence of the variance on \(\sigma^2\) are assumed for analytical convenience because in this way the prior is conjugate for \(\beta\). We also assume \(p(\sigma) \propto \sigma^{-1}\) as usual. The prior \(p(\alpha)\) is assumed to be flat, which would be common practice for both models (5) and (6) in the absence of time series considerations and specific prior information.

Let \(\bar{y}\) and \(\bar{x}\) be the sample means of \(y\) and \(x\), respectively, and let \(s_{yy}, s_{xx}\), and \(s_{xy}\) denote the sample variances of \(y\) and \(x\), and the sample covariance between \(x\) and \(y\), respectively. Given normality of the error term, the joint posterior of \((\alpha, \beta, \sigma)\) given the data \(D = (y, x, t)\) is written

\[ p(\alpha, \beta, \sigma | D) \propto \sigma^{-(T+2)} \left| \alpha \right|^{p/2} \exp \left[ -\frac{1}{2\sigma^2} (K |\alpha|^{p} \beta^2 + e' e) \right]. \]  

(8)

Standard integration formulas can be used to derive the marginal posterior of \(\alpha\). The general form of the marginal density of \(\alpha\) is given by

\[ p(\alpha | D) \propto g_T(\alpha)^{1/2} (Q^2(\alpha) + K g_T(\alpha) (\bar{y} - \alpha \bar{x}))^{-T/2}, \]  

(9)

where

\[ Q^2(\alpha) = s_x^2(\alpha - \hat{\alpha})^2 + s_e^2, \]

with \(\hat{\alpha} = s_{xy}/s_x^2, s_e^2 = s_y^2 - s_{xy}^2/s_x^2, g_T(\alpha) = (K + T |\alpha|^{2-p})^{-1}\).

Due to the nonlinearity, the exact marginal posterior of \(\beta\) cannot be obtained in a closed form for general \(p\) and \(K\). Exact analytical integration over \(\alpha\) is only possible for \(p = 0\). For \(p = 2\), the Laplace integration formula (see Phillips [2]), provides an approximation for the marginal posterior of \(\beta\). For \(p = 1\) and \(K > 0\), integration of (8) is complicated because of the nondiffer-
entiability of $|\alpha|$ at $\alpha = 0$; the Laplace formula does not yield a valid expression that holds uniformly in $\beta$. The results for various specific values of $p$ and $K$ are summarized in Table 1, together with the corresponding marginal posterior densities of $\alpha$. We briefly discuss the properties of these posteriors in a series of remarks below.

**Remarks.**

a. The behavior of the marginal posterior around $\alpha = 0$ is primarily determined by the factor $g_T(\alpha)$, which depends critically on the prior parameters $p$ and $K$. With $p < 2$, the weight factor $g_T(\alpha)$ attains its maximum at $\alpha = 0$. If $T \to \infty$, then $g_T(\alpha) \to 0$ at rate $1/T$, except at $\alpha = 0$, reflecting the local unidentifiability of $\beta$.

b. The "sum of squared residuals" $S^2(\beta)$ defined in Table 1 is a ratio of two quadratic functions in $\beta$, which converges to $s_0^2$ as $\beta \to \pm \infty$. The term $(z'z)$ is of the order $|\beta|^{-2}$ as $\beta \to \pm \infty$.

The technical results from Remarks a and b allow us to characterize the various marginal posteriors in Remarks c–h.

c. With $K = 0$, the marginal posterior of $\alpha$ is dominated by a factor $|\alpha|^{-(1-p/2)}$ close to $\alpha = 0$. This implies that flat independent priors on $\alpha$ and $\beta$ in the non-linear representation in (2), i.e., $p = 0$, lead to an improper posterior. The marginal posterior for $p = 1$ has an integrable singularity at $\alpha = 0$. With $p = 2$, the marginal posterior is Student-$t$.

d. With $K = 0$, the marginal posterior of $\beta$ has tails of order $|\beta|^{-(1+p/2)}$ as $|\beta| \to \infty$. This implies that the posterior is improper for $p = 0$. For $p = 1$ and $p = 2$, the marginal posterior is proper, but moments do not exist. Within the class of priors in (7), an improper prior on $\beta$, i.e., $K = 0$, never provides existence of posterior moments for any value of $p$.

e. With $K > 0$, the marginal posterior of $\alpha$ is always proper and without singularities at $\alpha = 0$. If $K$ is small and $p = 1$, the marginal posterior $p(\alpha|D)$ will often be bimodal with one mode close to $\alpha = 0$. For $p = 2$, the marginal posterior is exactly Student-$t$. Due to the factor $g_T(\alpha)$, the marginal posterior will be more concentrated around $\alpha = 0$ the smaller the value of $p$. In general, for given $K$ the posterior probability $Pr(|\alpha| > A)$ will be largest for $p = 2$ and smallest for $p = 0$.

f. With $K > 0$, the marginal posterior of $\beta$ is always proper. For $p = 0$, the tails are like a Student-$t$, a property it simply inherits from the prior. For $p = 2$, the tails are still of order $|\beta|^{-2}$, as in the diffuse case $K = 0$. In that respect, the case $p = 2$ is very different, because posterior moments of $\beta$ do not exist, even with a proper conditional prior.

g. With $K > 0$ and $p = 1$, the Laplace formula for integration over $\alpha$ is not uniformly applicable for all $\beta$ (Table 1). The posterior can also be characterized by using the conditional posterior moments,

$$E[\beta|\alpha, D] = \frac{T\alpha(y - \alpha \bar{x})}{K|\alpha| + T\alpha^2} \quad (10)$$
Table 1. Marginal posterior densities of $\alpha$ and $\beta$ in the model $y = \alpha(x + \beta) + \epsilon$

<table>
<thead>
<tr>
<th>$p = 0$ (prior independence $\alpha$ and $\beta$)</th>
<th>$p = 1$</th>
<th>$p = 2$ (prior independence $\alpha$ and $\delta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K &gt; 0$  $p(\alpha</td>
<td>D) \propto (K + \alpha^2 T)^{-1/2}$</td>
<td>$p(\alpha</td>
</tr>
<tr>
<td>$\times \left[ Q^2(\alpha) + \frac{K(y - \alpha \bar{x})^2}{K + T} \right]^{-T/2}$</td>
<td>$\times \left[ Q^2(\alpha) + \frac{K(y - \alpha \bar{x})^2}{K + T} \right]^{-T/2}$</td>
<td></td>
</tr>
<tr>
<td>$p(\beta</td>
<td>D) \propto (z'z)^{-1/2} \left[ S^2(\beta) + \frac{K}{T} \beta^2 \right]^{-T/2}$</td>
<td>$p(\beta</td>
</tr>
<tr>
<td>$K = 0$  $p(\alpha</td>
<td>D) \propto</td>
<td>\alpha</td>
</tr>
<tr>
<td>$p(\beta</td>
<td>D) \propto (z'z)^{-1/2} y'z S^2(\beta)^{-T/2}$</td>
<td>$p(\beta</td>
</tr>
</tbody>
</table>

$Q^2(\alpha) = s_y^2 + s_x^2 (\alpha - \bar{\alpha})^2$, with $s_y^2 = s_y^2 - \frac{s_y^2}{s_x^2}$, and $\bar{\alpha} = s_{xy}/s_x^2$.

$S^2(\beta) = \frac{1}{T} \left[ y'y - \frac{(y'z)^2}{z'z} \right]$, with $z = x + \beta$, so that $\frac{1}{T} z'z = s_x^2 + (\bar{x} + \beta)^2$, $\frac{1}{T} z'y = s_{xy} + \bar{y}(\bar{x} + \beta)$, $\frac{1}{T} y'y = s_y^2 + \bar{y}^2$.

$\psi(\beta) = + \frac{1}{2} K \beta^2$, if $\beta \in B_1 = \left\{ \beta : z'y + \frac{K}{2T} \beta^2 < 0 \right\}$,

$= - \frac{1}{2} K \beta^2$, if $\beta \in B_2 = \left\{ \beta : z'y - \frac{K}{2T} \beta^2 > 0 \right\}$.

The parameters $\alpha$, $\beta$, and $\delta$ are defined in equations (5) and (6) in the text; the prior is defined in (7). Marginal posteriors of $\alpha$ are obtained by exact analytical marginalization of the joint posterior $p(\alpha, \beta, \delta)$ in (8). The marginal posteriors of $\beta$ for the cases $p = 1, 2$ are obtained approximately by applying the Laplace integration formula (see Phillips [2] and Tierney, Kass, and Kadane [8]). For $p = 1$ and $K > 0$, the joint posterior in (8) is not differentiable at $\alpha = 0$. The reported marginal density of $\beta$ is given for the region where the "sum of squared residuals" is minimized at $\hat{\alpha}(\beta) = 0$. Solving the relevant inequalities gives the regions $B_1$ and $B_2$, which are simple functions of the sample moments and the prior precision $K$. The Laplace approximation does not apply at extreme values of $\beta$, and for the interval between $B_1$ and $B_2$, where $\hat{\alpha}(\beta) = 0$. Either $B_1$ or $B_2$ can be empty, and in fact will usually be for large $T/K$. 


\[ \text{Var}[\beta | \alpha, \mathbf{D}] = \frac{Q^2(\alpha)}{K|\alpha| + T\alpha^2}. \] (11)

The unconditional moments are found by integrating the conditional moments over \( \alpha \) by using the marginal posterior of \( \alpha \). The conditional mean is well behaved as \( \alpha \to 0 \), but the conditional variance is of order \( |\alpha|^{-1} \). Because the marginal posterior density \( p(\alpha | \mathbf{D}) \) is regular at \( \alpha = 0 \) and has tails like the Student-\( t \), the unconditional mean exists, but the variance does not exist.

h. If instead of the scalar parameter \( \beta \) we would have an \( (n \times 1) \) vector of parameters \( \beta \) with an associated \((T \times n)\) matrix of explanatory variables \( \mathbf{Z} \), all results with \( K > 0 \) remain valid. However, if \( K = 0 \), the marginal posterior of \( \alpha \) becomes of order \( |\alpha|^{-n(1-p/2)} \) close to \( \alpha = 0 \) and will be improper for all \( p < 2 - 1/n \). Hence, a flat prior on the linear representation is the only case for which the posterior is always proper, irrespective of the number of regressors \( n \).

The practical differences between the three priors \( p = (0, 1, 2) \) can be seen by comparing the marginal posteriors for a typical artificial data set. A sample is chosen such that the least-squares estimator gives \( \hat{\alpha} = 0.1 \) and \( \hat{\beta} = 6 \), the regression \( t \)-statistics are \( t(\hat{\alpha}) = 2 \) and (asymptotically) \( t(\hat{\beta}) = 4 \), whereas the \( R^2 \) of the regression equals 0.2, and \( \text{Var}(\varepsilon) = 1.2 \). The prior precision is \( K = 1 \). The sample moments and prior are such that the data are informative, but not overwhelmingly so. The relation between \( y \) and \( x \) is not entirely clear from the data; neither is the possibility that \( \hat{\beta} \) is a spurious estimate of the unconditional mean of \( x \). Figure 1 shows the marginal posteriors \( p(\alpha | \mathbf{D}) \) and \( p(\beta | \mathbf{D}) \). Table 2 reports the exact sample moments and some properties of the posteriors.

The figures and the table have a clear message. The \( p = 0 \) prior performs very poorly, even though it is the only prior for which the posterior mean and variance of \( \beta \) exist. The marginal posterior of \( \alpha \) is not at all concentrated around \( \alpha = 0.1 \), and the marginal posterior \( p(\beta | \mathbf{D}) \) is still centered around \( \beta = 0 \). With \( p = 0 \), the posterior means and modes are far from the classi-

---

**Table 2.** Posterior moments of \( \alpha \) and \( \beta^a \)

<table>
<thead>
<tr>
<th>( p )</th>
<th>( E[\alpha] )</th>
<th>S.D. [( \alpha )]</th>
<th>Mode[( \alpha )]</th>
<th>( E[\beta] )</th>
<th>S.D. [( \beta )]</th>
<th>Mode[( \beta )]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.003</td>
<td>0.031</td>
<td>0.000</td>
<td>0.38</td>
<td>1.10</td>
<td>0.18</td>
</tr>
<tr>
<td>1</td>
<td>0.048</td>
<td>0.056</td>
<td>0.000</td>
<td>4.03</td>
<td>–</td>
<td>5.99</td>
</tr>
<tr>
<td>2</td>
<td>0.091</td>
<td>0.061</td>
<td>0.091</td>
<td>–</td>
<td>–</td>
<td>6.59</td>
</tr>
</tbody>
</table>

\(^a\) \( E[\cdot] \) denotes the posterior mean, S.D. \([\cdot]\) the conditional standard deviation, and Mode\([\cdot]\) the posterior mode of the marginal density. The model is \( y = \alpha(x + \beta) + \varepsilon \), and the prior is \( \beta \sim N(0, \sigma^2 |\alpha|^{-\nu}/K) \) for \( p = (0, 1, 2) \) and \( K = 1 \). Sample moments of the data are \( \bar{x} = -5, \bar{y} = 0.1, \sigma_x^2 = 1.25, \sigma_y^2 = 5, s_{xy} = 0.5, \) and \( T = 64 \). The implied least-squares estimates are \( \hat{\alpha} = 0.1, \hat{\delta} = 0.6, \) and \( \hat{\beta} = \hat{\delta}/\hat{\alpha} = 6 \).
Figure 1. This figure shows the marginal posterior densities of $\alpha$ (A) and $\beta$ (B) in the model $y = \alpha (x + \beta) + e$ by using the prior $\beta \sim N(0, \sigma^2 |\alpha|^{-p} / K)$ for $p = 0, 1, 2$ and $K = 1$. Sample moments of the data are $\bar{x} = -5$, $\bar{y} = 0.1$, $s_y^2 = 1.25$, $s_x^2 = 5$, $s_{xy} = 0.5$, and $T = 64$.

...cal least-squares estimates, although the standard deviations of $\alpha$ and $\beta$ are small. The posterior mean of $\alpha$ increases with $p$, in line with remark e. The marginal posteriors of $\beta$ with $p = 1$ and $p = 2$ are very similar. The only difference is the tail behavior. The tails of the $p = 1$ density decline fast...
enough that the posterior mean of $\beta$ exists, which would make the $p = 1$ prior attractive.

The paradoxical conclusion of this section is that, to obtain a prior that leads to a well-behaved posterior mean of the parameter of interest $\beta$, we also end up with a marginal posterior of $\alpha$ that is more concentrated to the singularity at $\alpha = 0$ than the flat prior on $(\alpha, \delta)$ in the linear reduced form. But as the point $\alpha = 0$ causes the problems in determining the ratio $\beta = \delta/\alpha$, one would have expected a posterior that would downweight the probability that $\alpha$ is close to zero.

3. PRIORS FOR THE AR(1) MODEL WITH UNKNOWN MEAN

There are several ways to obtain the $p = 1$ type prior for the AR(1) model with unconditional mean $\mu$ defined in (1). Table 3 provides a list with priors suggested in the literature and the implied marginal posteriors of $\rho$. In this section, we will clarify the relations between the various priors and posteriors by drawing on the results of Section 2.

Schotman and van Dijk (SVD [4]) were largely motivated by the singularity discussed in Section 2, which they combined with time series considerations. The SVD prior is obtained from the initial condition of a stationary AR(1) and is given by $\mu | \rho, \sigma \sim N(\gamma_0, \sigma^2/(1 - \rho^2))$. Because in applications we are mostly interested in the area close to the unit root $\alpha = 1 - \rho = 0$, the variance in the conditional prior can be approximated by $\sigma^2/2\alpha$, which gives the case $(p = 1, K = 2)$. The $p = 1$ prior arises from the variance of the time series $\{y_t\}$ and not from the singularity itself.

Whereas the conditional prior for $\mu$ explicitly takes into account time series considerations, the flat prior on $\rho$ is arbitrary. The Jeffreys prior derived by Phillips [2], conditional on an initial observation $y_0 = \mu$, can for the structural parameters in model (2) be written as

$$p_J(\rho, \mu, \sigma) \propto \sigma^{-2} \phi(\rho)^{1/2} |1 - \rho|,$$

(12)

where $\phi(\rho) = (1 - \rho^2)^{-1}(T - (1 - \rho^{2T})/(1 - \rho^2))$. This prior has become known as the "critics" prior. Zivot [14] factorized and interpreted this prior as the product of $p(\mu | \rho, \sigma) \propto \sigma^{-1} |1 - \rho|$, $p(\sigma) \propto \sigma^{-1}$, and $p(\rho) \propto \phi(\rho)^{1/2}$, which would classify this prior as $(p = 2, K = 0)$. The form of the marginal prior $p(\rho)$ changes the behavior of the marginal posterior of $\rho$ such that it becomes very close to a prior with $p = 1$. For $\rho$ close to one, we approximate $1 - \rho^{2T}$ by the third-order Taylor expansion

$$1 - \rho^{2T} = 2T(1 - \rho) - T(2T - 1)(1 - \rho)^2 + \frac{3}{2} T(2T - 1)(T - 1)(1 - \rho)^3.$$

(13)

Substituting in the expression for $\phi(\rho)$ eventually gives
Table 3. Priors and marginal posteriors for the AR(1) model

| Reference                        | Prior \( p(\mu, \rho, \sigma) \) on structural parameters | Marginal posterior \( p(\rho | D) \) |
|----------------------------------|----------------------------------------------------------|-------------------------------------|
| DeJong and Whiteman [1]         | \( \sigma^{-1} |1 - \rho| \)                          | \( Q_0^2(\rho)^{-7/2} \)            |
| Phillips [2]                    | \( \sigma^{-2} \left[ \phi(\rho) + \frac{(y_0 - \mu)^2}{\sigma^2} \frac{1 - \rho^2 T}{1 - \rho^2} \right] |1 - \rho| \)                         | \( \phi(\rho)^{1/2} Q_0^2(\rho)^{-7/2} \) |
| Schotman and van Dijk [4]       | \( \sigma^{-2} (1 - \rho^2)^{1/2} \exp \left( -\frac{1 - \rho^2}{2\sigma^2} (y_0 - \mu)^2 \right) \) \( 1 + T \frac{1 - \rho}{1 + \rho} \) \( 1 / (1 + \rho) \) \( Q_1^2(\rho)^{-7/2} \) \( Q_1^2(\rho)^{(T+1)/2} \) |
| Uhlig [9], Zellner [11]         | \( (1 + T \frac{1 - \rho}{1 + \rho})^{1/2} \left( (T + 1)(1 - \rho^2) + 2\rho^2 \right)^{1/2} \) \( T + 1 + \frac{2\rho^2}{1 - \rho^2} \) \( Q_2^2(\rho)^{(T+1)/2} \) \( Q_2^2(\rho)^{-7/2} \) |
| Zivot [14]                      | \( \sigma^{-2} \exp \left( -\frac{(y_0 - \mu)^2}{2\sigma^2} \right) \) \( 1 + T(1 - \rho)^2 \)^{-1/2} \( Q_3^2(\rho)^{-7/2} \) \( Q_3^2(\rho)^{1/2} \) \( Q_3^2(\rho)^{-7/2} \) |
| Zellner [12]                    | \( \sigma^{-1} (1 - \rho^2)^{1/2} \) \( 1 + \rho \)^{-1/2} \( Q_4^2(\rho)^{-7/2} \) \( Q_4^2(\rho)^{1/2} \) \( Q_4^2(\rho)^{-7/2} \) |

\( Q_0^2(\rho) = s_0^2 + s^2_{x,-1} (\rho - \hat{\rho})^2 \), with \( s_0^2 = s^2_0 + s^2_{x,-1} / s^2_{x,-1} \), and \( \hat{\rho} = s_{x,-1} / s^2_{x,-1} \).

\( Q_1^2(\rho) = Q_0^2(\rho) + \frac{(\hat{y} - \rho \hat{y}_{-1} - (1 - \rho) y_0)^2}{1 + T(1 - \rho)^2} \).

\( Q_2^2(\rho) = Q_0^2(\rho) + \frac{(\hat{y} - \rho \hat{y}_{-1} - (1 - \rho) y_0)^2}{1 + T(1 - \rho)^2} \).

All priors are given in the form as would be applicable to the components representation in (2). This implies that we have multiplied by the Jacobian \( |1 - \rho| \) when the original prior was designed for the linear representation in (1), e.g., the DeJong and Whiteman [1] prior. For the Jeffreys prior derived from the exact likelihood function (Uhlig [9] and Zellner [11]), the prior is given as applicable after processing the initial observation. The marginal posterior of \( \rho \) for the exact likelihood Jeffreys prior and for the SVD [4] prior are only defined when \( |\rho| < 1 \).
\[ \phi(\rho) \approx 2T(T - 1) \frac{1 - (2T - 1)(1 - \rho)/3}{(1 + \rho)^2}, \] (14)

which close to \( \rho = 1 \) is well approximated by the simple formula

\[ \tilde{\phi}_c \propto (1 + cT(1 - \rho))^{-1}, \quad \text{with } \rho \leq 1 + (cT)^{-1} \] (15)

for \( c = 2/3 \). The factor \((1 + cT\alpha)^{-1/2}\) is of the same form as the factor \(g_T(\alpha)^{1/2}\), which appears in the marginal posterior derived from a \( p = 1 \) prior with \( K = 3/2 \). Over the stationary interval, the "critics" posterior is very similar to the posterior derived from the SVD prior. They share the factor \((1 + cT\alpha)^{-1/2}\), which pulls the marginal posterior away from the Student-\( t \) kernel toward the unit root. Because \( 3/2 < 2 \), the (truncated) "critics" prior leads to a posterior that is slightly more concentrated to the unit root. The effect of the term \( g_T(\alpha)(\tilde{y} - \alpha\bar{x})^2 \) in the SVD posterior is due to the initial condition and is negligible in large samples and for data sets where the sample means are small relative to the variance. This term appears in the Schotman–van Dijk posterior because of the prior on \( \mu \). The implications for \( \mu \) are different for the two priors. Because the "critics" prior is formally a \( p = 2 \) prior, the tails of the posterior on \( \mu \) are of the Cauchy type, whereas they are Student-\( t \) for the SVD prior.

Although derived from different motivations, the Phillips prior and the Schotman and van Dijk prior happen to lead to posteriors of \( \rho \) that are analytically very close. The similarity is not entirely coincidental because both priors involve the typical expression \((1 - \rho^2)^{-1}\) for the variance of an autoregressive time series. In the SVD prior, it enters through the initial condition, whereas for the "critics" prior, it comes in through the information matrix, which contains the term \( \mathbb{E}[\sum y_i^2 | y_0] \).

More priors of the \( p = 1 \) type can be derived by altering the assumptions on the model slightly, or by using different approximations of the priors discussed so far. For instance, if \( T \) is large and \( \rho < 1 \), the "critics" prior \( p(\rho, \mu, \sigma) \) can be approximated as being proportional to \( \sigma^{-2}(1 - \rho)/(1 + \rho)^{1/2} \). This prior is essentially proportional to \((1 - \rho)^{1/2}\), which leads to the classification \((p = 1, K = 0)\), i.e., the most diffuse prior in the \( p = 1 \) class. Hence, the marginal posterior of \( \rho \) will have a spike at the unit root (see Table 3) and will be more concentrated toward the unit root than the original "critics" prior. Another way to obtain the same prior is to reintroduce the initial condition in the Jeffreys prior of Phillips [2] (see also Schotman and van Dijk [5]),

\[ p(\rho, \mu, \sigma) \propto \sigma^{-2} \left[ \phi(\rho) + \frac{(y_0 - \mu)^2}{\sigma^2} \left( \frac{1 - \rho^{2T}}{1 - \rho^2} \right)^{1/2} \right] |1 - \rho|. \] (16)

Replacing \((y_0 - \mu)^2\) by its expectation conditional on \( \rho^2 < 1 \) again yields the factor \([((1 - \rho)/(1 + \rho))^{1/2}\). The same prior is also discussed in Zellner [12],
who derived it as the MDIP for the components representation (2). Zellner [11] interpreted this prior as a (proper) beta prior for $\rho < 1$.

Taking expectations over the initial condition $y_0$ requires an assumption about the start-up time of the process. The usual assumption is that the process started in the infinite past, in which case we must also assume stationarity to obtain the so-called exact likelihood. With the exact likelihood, the SVD prior is nothing but a flat prior on the structural parameters $\rho$ and $\mu$. The Jeffreys prior for this likelihood is derived in Zellner [11] and reproduced in Table 3. The most important part is the last factor in the prior, which involves $(1 - \rho^2)^{-1/2}$ and will dominate the posterior close to $\rho = 1$. The singularity is integrable, however, and of the same order as arises with the $(p = 1, K = 0)$ prior. The same factor occurs in the prior proposed by Uhlig [9], who extended the Jeffreys prior to deal with explosive models and a likelihood function for the initial observation. Because the Jeffreys prior does not functionally depend on $\mu$, the marginal posterior of $\rho$ follows by straightforward integration over $\sigma$ and $\mu$. It also implies that the posterior of $\mu$ will be fat-tailed.

The $p = 0$ and $p = 2$ priors have also been used recently. DeJong and Whiteman [1] employed a flat prior on the linear parameterization in (2), i.e., the case $(p = 2, K = 0)$. Given the properties of these priors, this implies that the resulting marginal posterior of $\rho$ will be more concentrated than the posteriors resulting from any of the above priors, and that most of the probability mass will usually be farther away from the unit root.

Because of a different treatment of the initial observation, Zivot [14] obtained a conditional prior for $\mu$ with $(p = 0, K = 1)$. Because this prior is of the $p = 0$ type, the simulation results reported by Zivot [14] are consistent with the results in Section 2, i.e., the posterior will be relatively more concentrated toward $\rho = 1$ than the “critics” posterior.

One way of examining the differences between the priors of DeJong and Whiteman [1], Schotman and van Dijk [4], and Zivot [14] is by the implicit assumption about the initial condition of the AR(1) process. Suppose we start out with a flat prior on $\mu$ and $\rho$ and the likelihood for the initial observation $y_0$, and suppose the resulting posterior is used as the prior for the rest of the sample. The three priors can be written in the form of the implicit assumption about the likelihood function for the initial observation:

$$p = 2: \quad y_0 = \delta + \rho y_0 + \nu_0(2), \quad \nu_0(2) \sim N \left(0, \frac{\sigma^2}{K} \right)$$

$$p = 1: \quad y_0 = \mu + \nu_0(1), \quad \nu_0(1) \sim N \left(0, \frac{\sigma^2}{1 - \rho^2} \right) \quad (17)$$

$$p = 0: \quad y_0 = \mu + \nu_0(0), \quad \nu_0(0) \sim N \left(0, \frac{\sigma^2}{K} \right),$$
where $\delta = \mu (1 - \rho)$. With $p = 2$, the initial condition $y_0$ is implicitly treated as a fixed point. As $K \to 0$, one obtains the DeJong and Whiteman [1] prior. The $p = 1$ prior has already been discussed at length in [4] and explicitly assumes a stationary process that started in the infinite past. The $p = 0$ prior with $K = 1$, as proposed by Zivot [14], avoids the strong stationarity assumption, but it is very tight around the initial condition. If $y_0$ is far from the sample mean, the posterior mean of $\mu$ can be very different from the classical ML estimate. On the other hand, loosening the prior by letting $K \to 0$ produces an improper posterior.

The posteriors that give relatively much weight to the unit root hypothesis, or more generally to values of $\rho$ close to one, have been evaluated positively because these priors somehow correct for the downward small sample bias of the classical $\hat{\rho}$ in the AR(1) model. Although this is a fortunate coincidence, and might well be a criterion in choosing a prior, it is not part of the principles by which any of the suggested priors was designed. Ex post, the Jeffreys prior turned out to be less biased than the flat prior, which added force to the arguments in favor of using it.

Because the small sample bias is important in the AR(1) model, it is interesting to repeat the example of Section 2 in a time series setting. The Nelson/Plosser data contain two series that do not show significant trend: the interest rate and unemployment. For the interest rate, an AR(1) model is often used in models for the term structure. The unconditional mean of the unemployment rate might be of interest as an indicator of a natural rate. Previous work with the extended Nelson/Plosser data (see Phillips [3] and Schotman and van Dijk [5]) showed that the interest rate is very close to a pure random walk, whereas the unemployment rate is almost surely a stationary series.

Figure 2 shows the marginal posteriors of $\mu$ and $\rho$ for the interest rate. Irrespective of the priors, the data recognize that this series is likely to be nonstationary. The Jeffreys prior for the exact likelihood is monotone, increasing toward its spike at the unit root. The $p = 0$ flat prior is nearly as sharply peaked toward the unit root. The "critics" prior and the SVD prior are almost indistinguishable and also have their mode at $\rho = 1$. The flat prior of DeJong and Whiteman [1] is the only one that leads to a posterior mode at $\rho < 1$. The ranking of the posteriors on the basis of the evidence in favor of the unit root is as we would expect from the analysis in Section 2, and the classification of the various priors given above.

Except for the $p = 0$ prior, the posteriors of $\mu$ are all very spread out, which reflects that the data contain very little information on the mean. The $p = 0$ prior for $\mu$ is a proper Student-$t$ density and centered around $y_0$; the posterior is still concentrated around the initial value and is not different from the prior. The posteriors on $\mu$ provide additional evidence on the plausibility of a unit root in the interest-rate series.

The posteriors for the unemployment rate in Figure 3 are very different.
Figure 2. Interest rate (1900–1988). This figure shows the marginal posterior densities of the unconditional mean $\mu$ (A) and the first-order autoregressive parameter $\rho$ (B) for the extended Nelson/Plosser series “interest rate” described in Schotman and van Dijk [5]. The graphs show posteriors for five different priors: a flat prior on the linear model in (1) as in DeJong and Whiteman (DJW [1]), the Jeffreys prior of Phillips (CRITICS [2]), the conditional normal prior of Schotman and van Dijk (SVD [4]), the Jeffreys prior for the exact likelihood function (EXACT), and an independent normal prior on the mean $\mu$ as in Zivot (FLAT [14]). The marginal posteriors of $\mu$ are obtained by numerical integration of the bivariate posterior of $\mu$ and $\rho$ over the stationary interval $|\rho| < 1$. 
**Figure 3.** Unemployment (1890–1988). This figure shows the marginal posterior densities of the unconditional mean $\mu$ (A) and the first-order autoregressive parameter $\rho$ (B) for the extended Nelson/Plosser series "unemployment" described in Schotman and van Dijk [5]. The graphs show posteriors for five different priors: a flat prior on the linear model in (1) as in DeJong and Whiteman (DJW [1]), the Jeffreys prior of Phillips (CRITICS [2]), the conditional normal prior of Schotman and van Dijk (SVD [4]), the Jeffreys prior for the exact likelihood function (EXACT), and an independent normal prior on the mean $\mu$ as in Zivot (FLAT [14]). The marginal posteriors of $\mu$ are obtained by numerical integration of the bivariate posterior of $\mu$ and $\rho$ over the stationary interval $|\rho| < 1$. 
Here the data are highly informative about the mean of the series. All priors lead to almost identical marginal posteriors on \( \mu \). The spike at \( \rho = 1 \) for the "exact" prior is now negligible because the cumulative probability mass close to \( \rho = 1 \) is less than 0.001. With the unemployment series, the data are sufficiently informative and far removed from the unit root, so that the differences between the priors is of no importance.

4. CONCLUSIONS

The identification problem of the unconditional mean in an AR(1) model under the unit root hypothesis affects the posterior of the autoregressive parameter. The weaker the prior dependence between these two parameters, the more the posterior of the autoregressive parameter will be shifted toward the unit root. The properties of different prior specifications have been investigated in a simple fixed regressor model. Time series considerations generally indicated a specific form of the prior dependence between the mean and the autocorrelation coefficient. An example with some of the Nelson/Plosser time series shows considerable sensitivity of the posterior with respect to the form of prior dependence if the data evidence on the unit root is weak.

The class of priors proposed in Section 2, with \( p = 1 \) and \( K > 0 \), is easily generalized to multivariate regression models, it leads to easily computable posteriors, the properties of the resulting posteriors are well understood, and the prior expresses the structural interpretation that we have by preferring a particular parameterization. Because the same singularity arises in higher-order autoregressive models, and also in an important model like the ECM, this type of prior might be applicable as an easy-to-use reference prior.

NOTES

1. To focus on the singularity, we keep the specification of the regression model as simple as possible. For notational convenience we assume that the explanatory variable associated with \( \beta \) is a constant term instead of some general explanatory variable \( z \). The assumption that \( \beta \) is a scalar, instead of the general case \( \beta'X \) with \( \beta \) an \( (n \times 1) \) vector, does have some consequences that will be discussed later in this section.

2. This was pointed out to me by a referee. The vector specification introduces \( n \) singularities because none of the elements in \( \beta \) will be identified at \( \alpha = 0 \). In the application to higher AR models there will, however, always be only one singularity because the higher-order AR terms enter additively in (2).

3. The approximation is very close over the range of positive stationary values \( 0 < \rho < 1 \). The prior is also defined for slightly explosive models, although the approximation breaks down rapidly for \( \rho > 1 \). The constant \( c = 2/3 \) comes from a local approximation around \( \rho = 1 \). The best approximating prior of the form \( (1 + cT(1 - \rho))^{-1} \) over the entire positive stationary interval turns out to be somewhat smaller. The informational difference \( \int_0^1 | \ln(\hat{\phi}(\rho)/\phi(\rho))| \phi(\rho) \, d\rho \) is (numerically) minimized for \( c = 0.6 \) for sample sizes ranging from \( T = 250 \) to \( T = 2,000 \).

4. The actual prior given by Zellner [12] has \( \sigma^{-1} \) instead of \( \sigma^{-2} \). Because the MDIP depends on the parameterization of the model, the MDIP for (2) differs from the MDIP for (1); see the discussion of the prior in Section 2.
5. However, see Uhlig [10] for a generalization to a startup time at some \( t = -S \), which allows nonstationary processes as well. Uhlig [10] takes the limit as \( S \) goes to infinity and explores its implications for Bayesian analysis.

REFERENCES


