Prior for Growth Rates, Small Sample Bias and the Effects of Monetary Policy*

Marek Jarociński
UPF, Barcelona and CASE, Warsaw

Albert Marcet
Institut d’Anàlisi Econòmica CSIC, UPF and CEPR

This version: March 17, 2006

Abstract
A Bayesian with a flat prior would say that OLS is the optimal estimator in a VAR. However, from a classical point of view, the well-known bias of OLS should be corrected. We show that OLS appears optimal to a Bayesian because the flat prior puts a very large weight on parameters that imply huge growth rates in the first few periods of the sample. This leads us to propose a prior that excludes huge growth rates. We argue that this prior, which is readily acceptable to most economists, breaks the classical/Bayesian dichotomy, as it can be interpreted as delivering a bias correction, and it has several advantages over proposed corrections for the bias. We show how to compute the posterior exactly and with a practical shortcut. We illustrate its effect in a VAR of Christiano et al. (1999), and show that the small sample bias causes OLS to understate the impact of monetary shocks on output.

Keywords: Small-Sample Bias, Bias Correction, Bayesian Methods, Vector Autoregression, Monetary Policy Shocks

JEL codes: E52, C11, C32

*Marek Jarociński thanks CEFIR, Moscow for hospitality during part of this work. Albert Marcet acknowledges financial support from CREI, DGES (Ministerio de Educación y Ciencia), CIRIT (Generalitat de Catalunya) and the Barcelona Economics program of CREA.
Contents

1 Introduction 3

2 Why OLS appear to be so good 6

3 The helicopter tour with a constant term 8
   3.1 Model without constant term, with flat prior . . . . . . . . . . 8
   3.2 Model with constant term and flat prior . . . . . . . . . . . . 10
   3.3 Introducing a prior on the initial growth rate . . . . . . . . . . 11

4 Estimation with the delta prior 14
   4.1 The delta posterior . . . . . . . . . . . . . . . . . . . . . . . 15
   4.2 The conditional posterior . . . . . . . . . . . . . . . . . . . . 19

5 Delta prior and classical bias corrections 21
   5.1 Modelling initial conditions . . . . . . . . . . . . . . . . . . 22
   5.2 Bayesian Bias Correction . . . . . . . . . . . . . . . . . . . . 24

6 Other Bayesian priors 27
   6.1 Minnesota, Jeffrey’s and shrinkage priors . . . . . . . . . . . . 27
   6.2 Priors similar to the delta prior . . . . . . . . . . . . . . . . . 28

7 Christiano et al. (1999) estimation of the effect of monetary
   shocks in the US 30
   7.1 Impulse responses to monetary shocks . . . . . . . . . . . . . . 30
   7.2 Comparison with Kilian (1998) bootstrap-after-bootstrap . . . 33

8 Conclusions 35

A S-priors for the AR(1) 38

B The S-prior for VAR’s 43

C Integrating constant for a multivariate normal likelihood 45

All the results were generated using Ox version 3.40 (see Doornik, 2002).
1 Introduction

The good finite-sample properties of OLS (unbiasedness and minimum MSE) only hold under the assumption of exogenous regressors. In particular, in a time series context, it is well known that OLS is biased and that it underestimates the highest root of the process for the relevant case of a root close to 1.\(^1\) Related to the bias is the well known fact that the short sample distribution of OLS is asymmetric, so that the standard methods for constructing confidence intervals based on symmetry are likely to be inaccurate. Most applied work on time series econometrics relies on asymptotics in order to justify using OLS and traditional (symmetric) confidence intervals, but this is just a way of ignoring the finite sample distribution.

The bias is substantial in VAR applications. The impulse-response coefficients are proportional to the root raised to a power, so that a small deviation in the VAR coefficients is magnified in the IRF coefficients at medium lags. The short-sample bias is likely to underestimate the size of the impulse responses and to diminish the actual significance of the results obtained.

This suggests that the OLS estimated root should be adjusted upwards to compensate for this bias. Some procedures have been designed for this purpose from a classical point of view.\(^2\) These procedures have been shown to reduce the MSE (in short samples) in some cases.

From a Bayesian point of view, however, using OLS and symmetric confidence intervals is easy to justify. The posterior distribution of the VAR coefficients under a flat prior is symmetric, centered at the OLS estimate. This implies that OLS is the optimal point estimate with a mean square error loss function. Many VAR applications nowadays are proceeding in this way.

Sims and Uhlig (1991) show that it is possible for the posterior distribution of parameters to be symmetric around OLS in spite of the skewness of the short sample distribution of the OLS estimator, and they conclude that using OLS is justified despite the short sample bias. Furthermore, these authors argue that the use of Bayesian estimators is often much easier than the use of classical finite sample corrections, so that they propose analyzing

---

\(^1\)The bias of OLS has been known and studied from a classical point of view since at least Quenouille (1949), Hurwicz (1950), Marriott and Pope (1954) and Kendall (1954).

VAR from a Bayesian point of view even if the researcher does not want to use any a-priori knowledge and, therefore, wishes to use a flat prior.

But we find this conclusion unsatisfactory. It appears there is a dichotomy between the Bayesian and classical approaches in time series. Embracing Bayesianism (and a flat prior) justifies using OLS, but most researchers would like to know why is it that OLS suddenly becomes a good estimator when you change your point of view. The short sample bias still is there and, therefore, it seems as if the (highest root of) OLS should be adjusted upwards.

In this paper we attempt to resolve these issues. First, we argue that the Bayesian and classical approaches are easily reconciled. We show that imposing a prior distribution that growth rates are unlikely to be very high (we call this a \textit{delta} prior) implies that the Bayesian posterior is asymmetric and it calls for an upward adjustment of OLS. We argue that the delta prior is not controversial to most researchers for many series of interest where the short sample bias would be an issue, so that it should be widely acceptable. We show how to compute the posterior exactly and with a very practical shortcut. As a result, using the delta prior can be understood as a way of correcting for the short sample bias in a practical way.\(^3\)

To make these points we first argue that the optimal properties of OLS hinge on the fact that the flat prior gives a very large weight to parameter values that imply huge growth rates at the beginning of the sample. Most economists will agree that the US GNP is unlikely to grow by \textit{more} than, say, 100\% in a quarter, but it is precisely the flat prior gives a very high probability to this event that OLS will appear to be a good estimator. We make this point in detail by performing a 'helicopter tour' similar to that of Sims and Uhlig (1991), and showing that the posterior distribution becomes asymmetric as soon as huge growth rates are given low probability. Then, OLS is no longer the estimator that minimizes mean square error to a Bayesian. A Bayesian would adjust OLS in the same direction as a classical econometrician, and the posterior probability intervals are asymmetric as in the distribution of OLS in short samples.

This leads us to propose a prior distribution that gives low probability to crazy initial growth rates. This prior is, we believe, widely acceptable: most economists will agree that US GNP is very unlikely to grow by more

\(^3\)The relation is, actually, very close: classical results consider initial values for the series that are not too far from its mean (or some 'counterpart' for the explosive processes), and such assumption has the effect of restricting the initial growth rate to be "reasonable".
than 100% in a quarter. The prior excludes precisely those parameter values that influence heavily the optimality of OLS under the flat prior and it links classical and Bayesian literature.

We show that, in fact, the delta estimator (i.e., the mean of the posterior under the delta prior) has very desirable properties even from a classical point of view, since it has a lower MSE than classical estimators, even those designed to improve short sample properties by correcting for the short sample bias.

We show how to implement this prior on growth rates with a shortcut that gives an approximation to posterior probabilities. Under this shortcut computational costs similar to those involved in computing the posterior in the standard case.

In addition, we also show a more time-consuming strategy to compute posterior probabilities exactly. This is of independent interest, as it shows how to compute posterior probabilities when the prior is over the behavior of the process itself and not on the parameters of the model. We show that the posterior has to be adjusted by 1) enlarging the actual sample with realizations of the model, 2) finding the maximum likelihood estimator for each enlarged sample, 3) averaging out over estimators obtained, giving small posterior probability to enlarged samples with low probability realizations and giving large posterior probability to samples that deliver an estimator close to the one with the original sample.

As an application, we show the impact of bias, classical bias corrections, and the delta prior on the estimates of the macroeconomic effects of monetary shocks. We concentrate on the well known VAR study of the US data, by Christiano et al. (1999). We first show that the effect of the small sample bias is readily visible in their reported results. We show that the estimation reported in CEE understates the effect of monetary shocks on output by almost a factor of two. It also somewhat underestimates the uncertainty about the obtained impulse responses, so that the price puzzle, (which CEE claim to have ruled out) is not completely rejected in our results.

The rest of the paper is organized as follows. Section 2 sets ground for our subsequent arguments by illustrating how initial conditions affect the precision of the OLS estimator in time series. Section 3 extends the Sims and Uhlig (1991) ‘helicopter tour’ including a constant term and it shows that adding the delta prior both the posterior and the short-sample distribution of OLS are asymmetric. Section 4 presents the delta prior and the computation of the posterior first under a convenient shortcut and second exactly. Section
5 explains the link between the delta prior and the classical time series bias literature, and it relates our prior to other priors for VARs proposed in the literature, including the Minnesota prior and the no-change prior of Sims (1996). Section 6 shows the implications of the prior on growth rates that we propose in Christiano et al. (1999)’s study of the effects of a monetary policy shock on macroeconomic variables in the US. We conclude in section 8.

2 Why OLS appear to be so good

We argue that a Bayesian is likely to exaggerate the virtues of OLS if parameter values that imply huge initial growth rates are given sufficiently high prior probability.

Precision of the OLS estimate depends on the variance of explanatory variables in the analyzed sample. In terms of undergraduate econometrics, the larger the dispersion of the explanatory variable, the smaller the term \((X'X)^{-1}\), and the lower the variance of OLS.

The dispersion, in the time series context, depends on the initial conditions. To illustrate this point, consider the AR(1) model with the intercept:

\[
y_t = \alpha + \rho y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \quad y_0 \text{ given (1)}
\]

\(\epsilon\) i.i.d. For the present example, we take: \(y_0 = 0, \rho = 0.95, \sigma = 1\) and the number of observations, \(T = 50\). The process has a stationary distribution with mean \(\mu = \alpha/(1 - \rho)\). Figure 1 displays two realizations of the process \(y\). The first column of graphs is a scatterplot of the above regression, each dot represents the independent variable \((y_{t-1})\) against the dependent variable \((y_t)\). The next column plots \(y_t\), and the third column plots \(\Delta y_t\), against time. The two panels only differ in the constant term \(\alpha\). Panel A presents a process with \(\alpha = 0\), while panel B presents a process with \(\alpha = 2\), which implies that it starts 40 standard errors away from the mean \((\mu = 40)\). The random errors \(\epsilon\) are the same in both panels. The solid line in the scatter-plots is the actual regression line implied by the true parameters in (1), the dashed line is the fitted regression estimated by OLS for this realization.

In the upper panel, the series fluctuates about its long run mean and the fitted regression line is significantly flatter than the truth. In this panel the realization contains small growth rates. Note that the slope of the dashed line in the scatter-plot is the value of the estimated OLS coefficient, and that
it is lower than the actual regression line, reflecting the fact that the OLS bias is quite large in that realization. In panel B, the transition from the remote starting value to the steady state, which dominates the first half of the sample, results in much higher dispersion of the values of the explanatory variable $y_{t-1}$. With this high dispersion of the explanatory variable OLS estimates the parameters much more precisely. This is why the fitted regression in panel B is much closer to the truth. Notice from the plots of the first difference of the series that $y$ grows unusually fast in the first periods, compared to the long-run growth rate.

Figure 1: Two cases of the AR(1) process and the performance of the OLS estimator of the coefficients. The first column shows scatter plots of $y_t$ against $y_{t-1}$ and the true and fitted regression lines. The second column plots the same $y_t$ against time. The third column plots first difference of $y_t$ against time.

It is well known that OLS is the best estimator (with a mean square error loss function) under a flat prior. But a flat prior on the constant term (given $y_0$) gives much more weight to parameter values that imply a mean far away from the starting point, as in panel B, than to parameter values.
that imply a mean close to the initial condition, as in panel A. OLS is very precise for parameter values as panel B, and no wonder that the Bayesian reasoning suggests it is optimal on average, since a flat prior gives a very large probability to realizations with very high initial growth rates.

We will argue that if we give low prior probability to cases like B, then the small sample bias of OLS becomes apparent also to a Bayesian, and the Bayesian posterior delivers a bias correction similar to the one in the classical literature.

3 The helicopter tour with a constant term

3.1 Model without constant term, with flat prior

For completeness we review here the results in Sims and Uhlig (1991) (SU). In the AR(1) model without constant term:

\[ y_t = \rho y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2) \]  

SU study the joint (bivariate) distribution of \( \rho \) and \( \hat{\rho} \). They point out that the result will be symmetric around a peak \( \hat{\rho} \).\(^4\) They find numerically this joint distribution and display different cuts of it in three-dimensional graphs.\(^5\)

The cross-sections along the fixed-\( \rho \) lines, as in Figure 2, represent the small sample distribution of \( \hat{\rho} \) given \( \rho \). The cross-sections along the fixed-\( \hat{\rho} \) lines as in Figure 3 represent the distribution of \( \rho \) given a value for the estimate \( \hat{\rho} \); these cross sections are a summary of the posterior distribution of

\(^4\)They point out that under the flat prior for \( \rho \), its posterior (assuming known \( \sigma \)), is gaussian:

\[ p(\rho|y) = N\left(\hat{\rho}, \frac{\sigma^2}{\sum_{t=1}^{T} y_t^2}\right) \]  

where \( \hat{\rho} \) is the OLS estimate. The variance of this distribution depends on the particular realized vector \( y \). To obtain the joint p.d.f. of \( \rho \) and \( \hat{\rho} \) one has to integrate on each instance of \( y \). But given that the posterior is symmetric, integrating \( y \) out will not destroy this symmetry (Sims and Uhlig, 1991, p.1593). As explained in Sims (1988), the downward bias of the OLS estimator is exactly compensated by the increasing precision of OLS for higher values of \( \rho \) and, as a result, the posterior is symmetric around the OLS estimator.

\(^5\)We have reproduced the numerical calculations of SU. We take a grid of values for \( \rho \) and, for each \( \rho \), generate 10000 data vectors of length \( T=100 \) from model (2), starting from \( y_0 = 0 \). For each data vector compute \( \hat{\rho} \). Lining up the histograms of \( \hat{\rho}'s \) obtain a surface which is the joint p.d.f. of \( \rho \) and \( \hat{\rho} \) under a flat prior for \( \rho \).
Figure 2: Figure 3 from Sims and Uhlig (1991): Joint frequency distribution of $\rho$ and $\hat{\rho}$ sliced along $\rho = 1$

Figure 3: Figure 4 from Sims and Uhlig (1991): Joint frequency distribution of $\rho$ and $\hat{\rho}$ sliced along $\hat{\rho} = 1$
$\rho$ given $\hat{\rho}$. These cuts reflect two well known facts: i) the short sample distribution of OLS under the classical approach (Figure 2) shows a downward bias of OLS as well as an asymmetric distribution, ii) the posterior under a flat prior represented by Figure 3 is symmetric and centered at the OLS estimator.

The difference in the two cuts justifies the dichotomy between the classical and Bayesian approach in time series models that SU advocated. If the researcher does not want to claim knowledge of the parameters, to the extent that the Bayesian estimator and posterior probabilities are optimally derived from Bayes’ rule, the relevant cut is the one in Figure 3, and this justifies, among other things, using plain OLS as a point estimate. A classical econometrician, however, would concentrate on Figure 2 and try to find a bias correction.

### 3.2 Model with constant term and flat prior

Key to our analysis of initial growth rates is the fact that the constant term is not known. Therefore, we first show that the results of SU’s helicopter tour extend to the case when the constant term in equation (1) is unknown and it needs to be estimated as well. In fact, the dichotomy between classical and Bayesian is deepened.

It has been pointed out from a classical point of view (for example, Andrews (1993, p.158)) that the small sample bias of OLS is much stronger when $\alpha$ has to be estimated. Therefore, it seems that from a classical point of view a correction for the bias is needed even more urgently in this case.

But a Bayesian is unimpressed by these classical results: it turns out that, under a flat prior, the posterior distribution in the model with a constant is also symmetric around OLS and, therefore, OLS remains the best estimator even if with the larger bias.\footnote{The reasoning about symmetry of the posterior $\rho|\hat{\rho}$ is similar as before. Now we have to consider the bivariate normal posterior for $(\rho, \alpha)$. We need to integrate out the $\alpha$, but the resulting marginal posterior of $\rho$ is:

$$p(\rho|y) = N\left(\hat{\rho}, \frac{\sigma^2}{\sum_{t=1}^{T} (y_{t-1} - \bar{y})^2}\right)$$

(4)

which is symmetric, and after integration w.r.t. $y$, the resulting posterior $\rho|\hat{\rho}$ distribution is also symmetric around $\hat{\rho}$.}
The joint bivariate distribution of \((\rho, \hat{\rho})\) is displayed in the next two figures.\(^7\) The cut along a given estimate is depicted in Figure 4 (which is the analog to figure 3). The distribution of \(\rho \mid \hat{\rho} = .95\) is indeed symmetric, the main change is that it is now very tightly concentrated along the \(\rho = \hat{\rho}\) line.

We do not show the cut for a fixed \(\rho\) to save space. The result is very similar to Figure 4.

This figure illustrates our claim that the dichotomy emphasized by SU is even stronger once the constant is estimated. Classical econometricians have emphasised that the small sample bias in a model with the constant is much stronger (Andrews, 1993). Therefore, upon observing \(\hat{\rho}\) close to 1, a classical econometrician would recognize that this estimate is a product of a large downward bias and believe that the true \(\rho\) is likely to be considerably greater. If this econometrician is concerned about short sample issues he would try to use a bias correction estimator. But, as follows from Figure 4, a Bayesian with a flat prior is even more convinced that the value of OLS is a good estimator of the true parameter when a constant is introduced. What is going on?

The simulation underlying Figure 4 makes it clear that the reason that the Bayesian views OLS as such a good estimator is that for stationary values of \(\rho (|\rho| < 1)\), large draws of the constant term correspond to large means of the process in the long run (given by \(\alpha/(1 - \rho)\)). Since the process always starts at \(y_0 = 0\), high probability of large \(\alpha\)'s means that realizations as in panel B of Figure 1 are very likely, in such cases OLS can be arbitrarily precise.\(^8\)

### 3.3 Introducing a prior on the initial growth rate

The discussion in the previous section leads us to introduce a prior on the growth rates. This gives low probability to observing realizations of the pro-

\(^7\)We just adapt the procedure of Sims and Uhlig (1991). We take the values of \(\rho\) on the grid; for convenience we assume a normal prior for \(\alpha \sim N(0, \sigma^2_\alpha)\) and draw constant terms \(\alpha\) from this distribution and take \(\sigma_\alpha \to \infty\). For each draw of \((\rho, \alpha)\) we simulate the process given by (1) starting from \(y_0 = 0\) and construct the bivariate distribution of \((\rho, \hat{\rho})\) following Sims and Uhlig (1991). \(\sigma_\alpha = 10\) turns out to be sufficiently big to make the cross-section along \(\hat{\rho} = 0.95\) close to a flat prior.

\(^8\)Large draws of \(\alpha\) unambiguously cause high growth rates in the beginning of the sample. \(\rho\) close to one also implies large mean of the process, but also a slower convergence to it, so its effect on the initial growth rate is ambiguous.
cess which start very far away from the mean, like in panel B of figure 1. We will see how the posterior becomes similar to the short-sample distribution.

The growth in period 1 is:

$$\Delta y_1 = \alpha + (\rho - 1)y_0 + \epsilon_1$$  \hspace{1cm} (5)

Our prior states that the growth attributable to the dynamics of the model (excluding the variability of $\epsilon_1$) is normally distributed with mean $g$ and variance $\sigma_g^2$:

$$\alpha + (\rho - 1)y_0 \sim N(g, \sigma_g^2)$$  \hspace{1cm} (6)

We call this the ‘delta’ prior.

Most economists will have an easy time specifying ‘reasonable’ prior mean $g$ and variance $\sigma_g^2$ for a given series. It is much easier to answer the question ‘what is a reasonable growth rate for output’ than to answer ‘what is a reasonable value for the roots of a process in a multivariate autoregression’.

A note of caution is needed about translating prior ideas on the growth rate into the delta prior. Notice that the total prior variance of the growth rate is given by

$$\text{Var}_p(\Delta y_1) = \sigma_g^2 + \sigma_\epsilon^2$$  \hspace{1cm} (7)

In words, the total a-priori uncertainty is given by the uncertainty about true parameters plus the underlying uncertainty of the shocks $\epsilon$. Hence, if an economist thinks that a reasonable prior standard deviation for the growth rate is 30%, this only gives the left side of (7), and the variance on the delta prior is then $\sigma_g^2 = (0.3)^2 - \sigma_\epsilon^2$.

Let us now call up a third helicopter and have a tour of the p.d.f. of $\hat{\rho}$ under this prior. In the present case, with $y_0 = 0$, the prior on growth rate just means $\alpha \sim N(g, \sigma_g^2)$, and the prior on growth rate simply translates as a prior on $\alpha$. We take $g = 0$ and $\sigma_g = 0.1$, which, if the data were in logs, would rule out with 95% probability initial 'systematic' (not caused by $\epsilon_1$) growth rates outside the (-20%,20%) range, which is a very loose requirement for most growing macroeconomic time series.

The resulting bivariate distribution of $(\rho, \hat{\rho})$ is depicted in figure 5, with a cut to display the distribution of $\rho \mid \hat{\rho} = .95$. Under the delta prior, the posterior is no longer symmetric around OLS: it is skewed towards larger values.\(^9\) There is no longer a dichotomy between Bayesian and classical views:

\(^9\)The reader should take notice that, following SU, the values on the $\rho$ axis values are increasing from right to left in the graph. This may be confusing at first, since it is natural
Figure 4: Joint frequency distribution of \((\rho, \hat{\rho})\) sliced along the \(\hat{\rho} = 0.95\) line (as in Figure 5 from Sims and Uhlig (1991)). Model with a constant term, and an approximately flat prior, starting from \(y_0 = 0\)

Figure 5: Joint frequency distribution of \((\rho, \hat{\rho})\) sliced along the \(\hat{\rho} = 0.95\) line (as in Figure 5 from Sims and Uhlig (1991)). Model with a constant term, drawn with \(\sigma_g = 0.1\), starting from \(y_0 = 0\)
the mean of Bayesian posterior would now adjust upwards the estimate, correcting for the OLS bias in a similar way as a classical econometrician would.

Therefore, if the huge initial growth rates as in panel B of Figure 1 are given small prior probability, a Bayesian ‘adjusts’ OLS in a similar way as a classical econometrician. This is not a coincidence. In most classical studies of the short-sample distribution of OLS the initial value of the variable is forced to be not too far from the mean of the process, so that the classical results have been implicitly using a prior similar to the one we propose. We study this in detail in section 5.

4 Estimation with the delta prior

In many practical applications, especially multivariate, introducing a prior on just the growth rate of the first period is not enough to influence the results. The large number of right hand side variables and parameters implies that just restricting the first growth rate does not have much of an impact on the posterior. For this reason, and because it is still (we believe) an acceptable prior to most observers of economic activity, we propose a prior that imposes a distribution on the growth rates in the first $T_0$ periods, where $T_0$ is fixed and does not grow with the sample size.

One may be tempted to impose the prior on all the observations in the sample and to set $T_0 = T$. But this is not reasonable, because it implies that the weight of the prior increases as the sample size grows, therefore the tightness of the prior grows with the sample size. Furthermore, imposing the delta prior on a large number of observations basically forces the parameter values to be consistent with a unit root which is, to our taste, too restrictive, and it goes away from the purpose of imposing minimal prior knowledge on the parameters and let the data speak as much as possible. Furthermore, the researcher may be willing to leave open the possibility that the initial growth rates are not similar to the long run growth rates, for example, in order to allow for a transition period due to an initial condition away from the steady state.

\footnote{to anticipate the upward correction to be illustrated by a shift of the probability mass to the right on the picture. We adopt this plotting convention for consistency with SU. They use it because it allows to switch smoothly between the two cross-sections of the plot in the ‘helicopter tour’ fashion, without having to flip one dimension.}
As usual, we model the process \( \{y_t\} \) as a \( J \)-dimensional VAR(\( P \)) process:

\[
y_t = \sum_{i=1}^{P} \Phi_i y_{t-i} + \Gamma' w_t + \epsilon_t \quad t \geq 0
\]

\( \epsilon_t \sim N(0, \Sigma_\epsilon) \) i.i.d. \( w_t \) is a vector of \( N \) exogenous variables - typically it would consist only of a 1, and then \( \Gamma \) would be a vector of constant terms. With \( T + P \) observations, we define \( Y \equiv [y_1, ..., y_T]' \) as a \( T \times J \) matrix gathering \( T \) observations determined by the model, and the VAR(\( P \)) process can be also written as:

\[
Y = XB + E
\]

Here, \( X \equiv [Y_{-1}, Y_{-2}, ..., Y_{-P}, W] \) collects the lagged values of \( Y \), and exogenous variables \( W \) (in the model with just the constant term, \( W \) contains only a column of 1’s), \( B \equiv [\Phi_1, ..., \Phi_P, \Gamma]' \) and \( E \equiv [\epsilon_1, ..., \epsilon_T]' \). The initial conditions \( Y_0 \equiv [y_{-P+1}, ..., y_0]' \) are fixed.

Now, the delta prior states that the parameter values have to be consistent with reasonable growth rates in the first few periods. Say that a researcher is willing to express his/her prior uncertainty about growth rates by stating that

\[
\Delta y_t \sim N(g, \Sigma_\Delta) \quad \text{for } t = 1, ..., T_0
\]

Here, \( g \) is the vector of prior means and \( \Sigma_\Delta \) is the vector of total prior variance (that is, in terms of our discussion of equation (7), this is equal to \( \sigma_g^2 + \sigma_\epsilon^2 \)). Note that, given the initial conditions \( Y_0 \), this gives a distribution for the first \( T_0 \) observations of the level of the series \( y \).

The presumption is that most economists can easily find reasonable values for \( g \), and even if they choose large values for the variance \( \Sigma_\Delta \), the prior will be sufficient to perform a bias-adjustment that, as we will argue in section 5, often performs even better than classical estimators.

### 4.1 The delta posterior

To characterize the posterior, note as usual that the likelihood conditional on initial \( P \) observations is proportional to:

\[
L(Y; B, Y_0) \propto |\Sigma_\epsilon|^{-\frac{T}{2}} \exp \left( -\frac{1}{2} \text{tr} \Sigma_\epsilon^{-1} E'E \right)
\]

It turns out that the delta posterior is obtained by averaging out normal distributions obtained by stacking the data and realizations of \( y \) drawn from
the delta prior. Formally, let $\varphi$ denote the density of $[y_1, ..., y_{T_0}]'$ consistent with the delta prior and the observed initial conditions $Y_0$; i.e., $\varphi$ is obtained from $y_t \mid y_{t-1}, ..., y_1 \sim N(g + y_{t-1}, \Sigma_{\Delta})$ for $t = 1, ..., T_0$. Let $\overline{Y} \equiv [\overline{y}_1, ..., \overline{y}_{T_0}]'$ denote a particular realization of the first $T_0$ observations drawn from this density. Let $\hat{B}(Y, \overline{Y})$ be the OLS estimator of $B$ obtained with the sample of size $T + T_0$ resulting from stacking the actual sample $Y$ and a given realization $\overline{Y}$. Let $\hat{\Sigma}(Y, \overline{Y})$ be the variance-covariance matrix of the OLS estimator of $B$ in the stacked sample. Assume the variance of the shocks $\Sigma_{\epsilon}$ is known. In this proposition, assume for simplicity that $T_0$ is larger than the number of parameters to estimate, formally, $T_0 > K \equiv PJ + N$.

**Proposition 1** The posterior implied by the delta prior satisfies

$$\text{Post}(B; Y) = \frac{1}{\mathcal{K}(Y)} \int_{R^T \times J} \frac{\mathcal{K}(Y, \overline{Y})}{\mathcal{K}(\overline{Y})} \Phi \left( B; \hat{B}(Y, \overline{Y}), \hat{\Sigma}(Y, \overline{Y}) \right) \varphi \left( Y \right) dY$$

(10)

where

- $\Phi(\cdot; \mu, \Sigma)$ is the normal density with mean $\mu$ and variance $\Sigma$
- $\mathcal{K}(\cdot)$ is the integrating constant that converts the likelihood of the corresponding sample into a pdf that integrates to one. More precisely

$$\mathcal{K}(Y) \equiv \int L(Y; \overline{B})d\overline{B} =$$

$$= (2\pi)^{-\frac{(T-K)J}{2}} |\Sigma_{\epsilon}|^{-\frac{T-K}{2}} |X'X|^{-\frac{J}{2}} \exp \left( -\frac{1}{2} \text{tr} \left( \hat{E}' \hat{E} \Sigma_{\epsilon}^{-1} \right) \right)$$

(11)

(see Appendix C) where $K = PJ + N$ is the number of parameters in each equation

- $\mathcal{K}$ is an integrating constant, to be computed separately, insuring that (10) integrates to one.

**Proof**

Throughout the proof and the paper, $L$ is the likelihood function under normal distribution of the errors.
Let \( f_B \) be the prior distribution of the parameters implied by the delta prior. To find this prior, first note that the joint prior density of \((B, \overline{y}_1, \ldots, \overline{y}_{T_0})\) given by the delta prior is given simply by

\[
p(B, \overline{y}_1, \ldots, \overline{y}_{T_0}) = p(B \mid \overline{y}_1, \ldots, \overline{y}_{T_0}) \cdot p(\overline{y}_1, \ldots, \overline{y}_{T_0})
\]

\[
= \frac{L(\overline{Y}; B, Y_0)}{\mathcal{K}(\overline{Y})} \varphi(\overline{Y})
\]

(that is, the ‘posterior’ of \( B \) conditional on \( \overline{Y} \), times the marginal of \( \overline{Y} \)). The first equality is simple probability rules, the second equality uses familiar formulae for posterior of \( B \) when no additional knowledge of the parameters is available, and \( \mathcal{K}(\overline{Y}) \) is the integrating constant

\[
\mathcal{K}(\overline{Y}) \equiv \int \mathcal{L}(\overline{Y}; B, Y_0) \, dB
\]

Therefore, the prior satisfies

\[
f_B(B) = \int_{R^{T_0 \times J}} \frac{L(\overline{Y}; B, Y_0)}{\mathcal{K}(\overline{Y})} \varphi(\overline{Y}) \, d\overline{Y}
\] (12)

The usual formula for the posterior now gives

\[
\text{Post}(B; Y, Y_0) = \frac{L(Y; B, Y_0)}{\tilde{\mathcal{K}(Y)}} \int_{R^{T_0 \times J}} \frac{L(\overline{Y}; B, Y_0)}{\mathcal{K}(\overline{Y})} \varphi(\overline{Y}) \, d\overline{Y}
\]

We have abused notation slightly, since the first function \( L \) in this formula has a higher dimensional argument \( Y \) than the second \( L \) so these are really two different functions and they can only be distinguished by the argument. \( \tilde{\mathcal{K}} \) is the integrating constant of the whole posterior, which we will determine later.

Now, moving the first \( L \) inside the integral, multiplying out the exponential terms in the likelihoods and including the appropriate constant of integration we have

\[
\text{Post}(B; Y, Y_0) = \frac{1}{\tilde{\mathcal{K}(Y)}} \int_{R^{T_0 \times J}} \frac{L(Y; B, Y_0) \cdot L(\overline{Y}; B, Y_0)}{\mathcal{K}(\overline{Y})} \varphi(\overline{Y}) \, d\overline{Y}
\]

\[
= \frac{1}{\tilde{\mathcal{K}(Y)}} \int_{R^{T_0 \times J}} \frac{\mathcal{K}(Y, \overline{Y}) \cdot L(Y; Y, B, Y_0)}{\tilde{\mathcal{K}(Y, Y)}} \, \mathcal{K}(Y, \overline{Y}) \varphi(\overline{Y}) \, d\overline{Y}
\]
where, for each realization $\overline{Y}$, the term $L(\overline{Y}, \overline{Y}; \overline{B}, Y_0)$ is the likelihood obtained by stacking the actual sample and each realization $\overline{Y}$. The usual argument says that the normalized likelihood $\frac{L(\overline{Y}, \overline{Y}; \overline{B}, Y_0)}{K(\overline{Y}, \overline{Y})}$ is equal to a normal distribution centered at OLS of the sample $(Y, \overline{Y})$ and with the OLS estimated variance-covariance matrix. Therefore, $\frac{L(\overline{Y}, \overline{Y}; \overline{B}, Y_0)}{K(\overline{Y}, \overline{Y})} = \Phi \left[ \overline{B}; \hat{B}(Y, \overline{Y}), \hat{\Sigma}(Y, Y) \right]$, and this gives the result.

Clearly, the formula for $\tilde{K}$ integrates this posterior:

$$\tilde{K}(Y) = \int_{R_{B, \hat{B}(Y, \overline{Y})}} \frac{K(Y, \overline{Y})}{K(\overline{Y})} \Phi \left[ \overline{B}; \hat{B}(Y, \overline{Y}), \hat{\Sigma}(Y, Y) \right] \varphi(\overline{Y}) \, d(\overline{Y}, \overline{B})$$

□

The formula (10) says that the posterior is a weighted average of normal distributions, each distribution found by adding a hypothetical realization $\overline{Y}$ to the actual data. The weight given to each hypothetical realization is $\frac{K(Y, \overline{Y})}{K(\overline{Y})} \varphi(\overline{Y})$.

This weight is readily interpreted. Obviously, the term $\varphi(\overline{Y})$ just gives more weight to a priori more likely realizations. The ratio of integrating constants is related to the comparison of a model which accounts for the whole stacked sample $(Y, \overline{Y})$ with one $\overline{B}$, against a model which used separate coefficients for $Y$ and for $\overline{Y}$. To justify this more precisely, note

$$\frac{K(Y, \overline{Y})}{K(\overline{Y})} = K(Y) \frac{K(Y, \overline{Y})}{K(\overline{Y})K(Y)} =$$

$$= \frac{K(Y)}{2\pi} \frac{\Sigma_{\epsilon}}{\frac{1}{2}} \left( \frac{\frac{1}{2}}{|X'X|} \frac{1}{\frac{1}{2}} \right)^{\frac{1}{2}} \left( \frac{X'X}{X'X + \frac{1}{2}} \right) \times \exp \left( -\frac{1}{2} \tr \left[ (RSS(Y, \overline{Y}) - RSS(Y) - RSS(\overline{Y})) \Sigma_{\epsilon}^{-1} \right] \right)$$

where the first equality is trivial, the second equality uses the formula (11) and simple algebra. Here, $X$ are the regressors in the observed sample, $\overline{X}$ the regressors in the hypothetical realization, and $RSS$ is the residual sum of squares for each sample where residuals are computed by OLS applied to each sample. The ratio after the first equality is a Bayes factor comparing a model for the whole stacked sample with one using separate coefficients.
for $Y$ and for $\bar{Y}$. This ratio re-weighs each realization of $\bar{Y}$ and it gives less weight to realizations such that

1. $\bar{Y}$ behaves very different from $Y$ in the sense that a Chow test of constancy of coefficients would be rejected

2. the regressors $\bar{X}$ are not very informative about individual coefficients

To argue 1), notice that the term in the exponent in (13) is proportional to a Chow test, comparing the sum of square residuals of the restricted and unrestricted regressions when the unrestricted allows a whole different set of coefficients for some of the observations.

To argue 2) notice that the drawn regressors $\bar{X}$ are not informative if they are highly collinear, or have little variability. Then the determinant $|\bar{X}'\bar{X}|$ is close to zero while the other determinants involved in (13) need not be zero.

To implement this posterior it is easy to first draw a $\bar{Y}$ from the distribution $\varphi$, then draw a parameter from the corresponding normal distribution and to weigh this by the ratio of integrating constants inside the integral. This should allow computation, first of all, of the constant $\tilde{K}$, or, when using (13), the Bayes factor $\frac{K(Y)}{K(\bar{Y})}$ which will be more conveniently scaled. Then, all quantities of interest can be obtained from the weighted sample.

4.2 The conditional posterior

We now suggest an approximate way of finding the posterior. The idea is to impose the prior on the first $T_0$ observations but conditioning each observation $t$ on the observed past $y$'s. We now extend the discussion for the one-variable one-lag case, from section 3.

Since the growth rate is given by

$$\Delta y_t = B'(y_{t-1}, \ldots, y_{t-P}, 1) - y_{t-1} + \epsilon_t$$

(terms in the parenthesis are stacked vertically) we have that this growth rate is influenced by the shock $\epsilon$ and the model determined growth, given by $B'(y_{t-1}, \ldots, y_{t-P}, 1) - y_{t-1}$.

The prior determines the mean and variance for the model determined growth. Note that

$$E_{Pr} \Delta y_t = E_{Pr} [B'(y_{t-1}, \ldots, y_{t-P}, 1) - y_{t-1}] = g$$

19
\[ E_{P_t}[\Delta y_t \Delta y'_t] = \Sigma_\Delta = \text{var-cov}_{P_t}[B'(y_{t-1}, \ldots, y_{t-P}, 1) - y_{t-1}] + \Sigma_\epsilon \quad t = 1, \ldots, T_0 \]

therefore, the 'model determined' growth rates in the first \( T_0 \) observations are normally distributed with means \( g \) and variance

\[ \text{var-cov}[B'(y_{t-1}, \ldots, y_{t-P}, 1) - y_{t-1}] \equiv \Sigma_g = \Sigma_\Delta - \Sigma_\epsilon \quad (15) \]

and with the autocorrelation matrix \( S \) which is the same for all variables.

Normally we would have to integrate past \( y \)'s out of the above mean and variance, because the delta prior has implications for all \( y \)'s within dates 1 to \( T_0 \), but the shortcut we propose is to use the above formula by fixing past \( y \)'s that appear in (14) and (15) and setting them equal to the observed ones in the data. This gives

\[ \text{vec} \left( X_0 B - Y_{-1} \right) \sim N(g \otimes \iota_{T_0}, \Sigma_g \otimes S) \quad (16) \]

(where \( \iota_{T_0} \) denotes a column vector 1's, of length \( T_0 \) ) or

\[ p(\text{vec} \, B \, | \, X_0) \propto \exp \left( -\frac{1}{2} \left( \text{vec} \, B'(\Sigma_g^{-1} \otimes X_0'S^{-1}X_0) \, \text{vec} \, B - 2 \, \text{vec} \, B' \, \text{vec} \left( X_0'S^{-1}(Y_{-1} + \iota_{T_0}g')\Sigma_g^{-1} \right) \right) \right) \]

(17)

which is a normal prior, conjugate with the likelihood. If we use the standard RATS prior for the error variance:

\[ p(\Sigma_\epsilon) \propto |\Sigma_\epsilon|^{-\frac{v+J+1}{2}} \quad (18) \]

where \( v = PJ + K \), i.e. the number of variables in each equation, the posterior for \( B \) and \( \Sigma \) can be easily simulated with the Gibbs sampler. We will refer to this prior as the conditional delta-prior.

This prior is only 100% consistent with the delta prior in the case that \( T_0 = 1 \), because in that case the values of \( y \) in (14) and (15) are given in the initial conditions \( Y_0 \). But if the prior is imposed on more observations (i.e., if \( T_0 > 1 \)) past values that appear in the mean and variance are unknown a priori, they should be agreeable with the distribution implied by the prior and not set equal to the observed data. But this delta prior conditional on actual data is easy to use, it involves similar computations as the RATS prior and it is the one we use in this version of the working paper.
5 Delta prior and classical bias corrections

We argue in this section that the delta-prior, although inspired by Bayesian principles, can be also justified from a classical perspective. In the related paper (Jarociński and Marcet, 2005) we discuss at length the relationship between classical bias correction and Bayesian estimators. We argue there, that focusing on the bias is unjustified and that a classical estimator could be constructed that takes care of mean square error. We relate carefully Bayesian estimation and classical bias corrections by showing a series of estimators designed to behave well in short samples; each of these estimators is related to its predecessor, and the first estimator is close to bias correction while the last is close to Bayesian estimator. In this way, Bayesian and classical estimation can be regarded as being quite close.

In this section we focus on the properties of the delta prior relative to bias correcting estimators. The section is quite involved, so here we summarize the results. First we argue that studies of classical bias correction are related to the delta prior in two ways: these studies have specified certain assumptions for the distribution of the initial variable (initial conditions). Even though this was done for technical reasons (in order to suppress dependence on nuisance parameters), the assumptions for initial conditions that have been used turn out to avoid huge growth rates in the first few periods. In this sense, the literature on classical bias correction has imposed a requirement on initial conditions that works similarly to the delta prior. The second point of contact is that the posterior under the delta prior adjusts the mean of the estimator in the same qualitative direction as bias correction estimators and, in this sense, the delta prior could be described as providing a ‘Bayesian bias correction’.

But the bias corrections have two problems: first they actually give biased estimators, second, there is no reason that removing the bias is a good objective per se. It is well known that an unbiased estimator may have a very large mean square deviation. We show that the delta estimator is substantially better in terms of MSE than classical even when, in a classical spirit, we consider many fixed parameter values. The lower mean square error of the delta estimator obtains for a wide range of parameter values that include the parameters of interest in a model that grows, even when the parameter is fixed. We conclude that the delta-prior can be motivated on classical grounds.
5.1 Modelling initial conditions

This subsection discusses various ways to model the initial condition that have been used in the literature. Although it is a fairly long digression, we need it in order to set common ground for the discussion of classical bias corrections and alternative Bayesian priors proposed in the literature.

The first step is to reparameterize the model as:

\[ y_t - \mu = \rho (y_{t-1} - \mu) + \epsilon_t \quad \text{for} \quad t = 1 \ldots T \]  

(19)

This parametrization is a special case of (1) for:

\[ \alpha = \mu (1 - \rho) \]  

(20)

The specified distribution for \( y_0 \) will be related to \( \mu \). Under (19) the process reverts to \( \mu \) if \(|\rho| < 1\), goes away from \( \mu \) if \(|\rho| > 1\), and it drops out when \( \rho = 1 \).

Classical estimators need to specify fully the distribution of the process in order to derive the short sample properties of the estimator. One could assume, for example, \( y_0 = \mu \), but this would be very restrictive since, in practice, there is no reason that the first period in the sample is exactly at the centering parameter \( \mu \). Usually a distribution of the initial variable \( y_0 \) is specified.

Since the focus is on estimating \( \rho \), it is convenient to specify a distribution that causes the OLS estimator to be independent from the ‘nuisance’ parameters of the model: \( \mu \) and \( \sigma \). In that way, the results on estimators of \( \rho \) are valid for a wide range of nuisance parameters. A general condition for this independence is

**Result 2** Assume the initial condition in model (19) is given by:

\[ y_0 = \mu + \sigma \psi \]  

(21)

---

\(^{10}\)One can not go the other way around: if \( \rho = 1 \) in (1) there is no \( \mu \) that satisfies the above equation and generates the same \( y \)'s. This will have implications for the practical application of this parametrization when the observed variable has a positive growth rate and is close to a unit root, because a trend will have to be introduced in this case to have a growing variable.

\(^{11}\)Zivot (1994) explains that the motivation for suppressing the constant term \( \alpha \) when \( \rho = 1 \) is that it changes interpretation in the unit root case, from determining the level of the process to determining its drift. When no trend is included for the general \( \rho \), it is not reasonable to allow for a drift in just one point of the parameter space: \( \rho = 1 \).
where $\psi$ is a random variable. Then, if $\psi$ independent of the shocks $\epsilon$ and its distribution is independent of $\mu$ and $\sigma$, the distribution of the OLS estimator of $\rho$ in (1) is independent of $\mu$ and $\sigma$.\textsuperscript{12}

All classical papers that we know of use a version of (21). Of course, each choice of $\psi$ will result in a different small sample distribution of $\hat{\rho}$ conditional on $\rho$. Let us now review various assumptions on $\psi$.

The case $y_0 = \mu$ discussed above corresponds to $\psi = 0$. Bhargava (1986) constructs invariant (to $\mu$ and $\sigma$) tests for unit roots taking

$$
\psi \sim N \left( 0, \frac{1}{1 - \rho^2} \right) \quad \text{when } |\rho| < 1
$$

$$
\psi \sim N (0, 1) \quad \text{otherwise}
$$

so that the independence of the distribution of $\psi$ on the nuisance parameters required in the proposition holds. This yields $y_0 \sim N \left( \mu, \frac{\sigma^2}{1 - \rho^2} \right)$ in the stationary case, so it is assumed that the process is in the stationary distribution (as if it had been running forever). It yields $y_0 \sim N (\mu, \sigma^2)$ in the non-stationary case, that is, it says that $y_0$ was generated from the model and the initial condition was brought back one period to set $y_{-1} = \mu$.

\textsuperscript{12}Similar results have been used in the literature. As can be seen, the proof is very simple, but we could not find a formal proof, so we offer it here for completeness. The proposition is very similar to the property of $\hat{\rho}$ discussed in Andrews (1993, Appendix A), which contains a verbal proof for $|\rho| \leq 1$ and a particular distribution for $\psi$, but allowing for a trend.

**Proof:** Define normalized errors: $u = \epsilon / \sigma$. (21) allows to write:

$$
y_t = \mu + \sigma \left( \sum_{i=1}^{t} \rho^{t-i} u_i + \rho^t \psi \right) = \mu + \sigma \tilde{y}_t
$$

where $\tilde{y}$ is the process with the zero mean, which would obtain from the same realization of errors, but rescaled to have a unit variance. Then it is a matter of simple algebra to show that:

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

$\square$

23
Andrews (1993) considers median unbiased estimators for \( \rho \) by taking a similar starting point for the stationary case and an arbitrary starting point (i.e. \( \psi \) is an arbitrary constant) for the unit root case (he does not consider explosive values of \( \rho \)).

MacKinnon and Smith (1998) take:

\[
y_0 \sim N(\mu, \sigma^2)
\]  

so that they take \( \psi \sim N(0, 1) \) for all \( \rho \).

Uhlig (1994) proposes a formulation which encompasses many cases: he assumes \( y - S = \mu \), where \( S \) is a given constant that the researcher has to specify. This provides an alternative \( \psi \).

To provide a common nomenclature for all these cases, we call Uhlig’s proposal the ‘\( S \)' model. Then the case \( y_0 = 0 \) becomes the ‘\( S = 0 \)' model, the assumption of MacKinnon and Smith (1998) as in (23) becomes the ‘\( S = 1 \)' model and we call the assumption of Bhargava (1986) in (22) the ‘\( S = \infty \)' model.

We point out here that there are many ways of choosing the distribution of initial variables. It would seem that these initial conditions could have very different implications, and so that MacKinnon and Smith (1998) and Andrews (1993) may be correcting completely different biases.

### 5.2 Bayesian Bias Correction

We can now compare the delta-prior to classical bias correction. A Bayesian posterior mean is guaranteed to have the lowest MSE on average over the parameter space given by the prior. But from a classical point of view this is of little interest, the question is how an estimator performs for a given value of the parameters. Let us dub the mean of the delta posterior as the *delta estimator*. We show the Bias and MSE properties of delta estimator compared to alternative estimators from a classical point of view, for given parameter values.

The bias of various estimators is examined in a Monte Carlo study in which, in order to highlight the small sample problems, we assume \( T = 25 \). We consider the AR(1) model for various \( \rho \)'s. We model the initial condition with \( S = 1 \).\(^\text{13}\) We consider delta prior with mean 0 and two

\(^{13}\)To compute the values in this and next figure we took 100,000 realizations of the process and values \( \rho = 0, 0.05, \ldots, 1.2 \).
standard deviations: 0.2 (denoted delta1) and 0.05 (denoted delta2).

The performance of the delta estimator is compared with that of OLS and a bootstrap-bias-corrected (CBC) estimator of MacKinnon and Smith (1998) (that is, the bias correction is derived with the actual model $S = 1$) in a Monte Carlo experiment. Kilian (1998) essentially uses CBC to improve the small sample properties of inference with VARs. Figure 6 shows the bias for each estimator and each possible true value of $\rho$.\(^{14}\) Obviously, the largest bias is with OLS. The CBC estimator is only approximately unbiased (see, Jarocinski and Marcet (2005) for a discussion) so the picture reflects that the bias is not zero but, as is well known, CBC reduces the bias considerably relative to OLS. We can see that for high values of $\rho$ the bias of the delta estimator is in between that of OLS and CBC.\(^{15}\)

Therefore, the Bayesian estimator can also achieve a bias correction. The correction is not as 'good' as the CBC estimator. But it is clear that achieving (or getting close to) unbiasedness is not an important goal per se. A researcher cares about obtaining estimates that are on average close to the true value, so what is relevant is the total mean square deviation from the true parameter and, as is well known, unbiased estimators can have large MSE. We now turn to studying the MSE but still remain in the classical setup, and study the MSE for fixed values of $\rho$.

Figure 7 is a more relevant measure of how big deviations are, as it reports the root MSE for the three estimators under consideration at various values of $\rho$. The figure shows that Bayesian estimators beats OLS and the CBC for sufficiently high values of $\rho$. This surely contains the relevant roots in practice for possibly non-stationary series. Actually, the delta estimator can be substantially better, for example, at $\rho = 1$ the RMSE under CBC is around 30% larger than with the delta2 estimator. This increase is very large, close to the one obtained if half of the sample is thrown away.

Notice that, in this case, all the cards seem to be stacked in favor of the CBC: the bias correction is designed assuming $S = 1$ and this is, in fact, the truth. In the real world this (or any other) assumption on the initial condition is quite unwarranted, so we would expect CBC to be at an even

---

\(^{14}\)We only report positive values of $\rho$. Since we are concerned with possibly non-stationary series we can ignore negative values of $\rho$. These would imply oscillations in the series, so they imply large changes from one period to the next and, clearly, the delta prior does not work very well when this is the truth.

\(^{15}\)We have done similar computations when the CBC adjustment is computed with $S = 0$ and $S = \infty$. The general picture does not change.
Figure 6: Bias of the OLS, CBC and two delta estimators in a Monte Carlo experiment, sample size: $T=25$.

Figure 7: RMSE of the OLS, CBC and two delta estimators in a Monte Carlo experiment, sample size: $T=25$. 
larger disadvantage in practice. We have computed all possible combinations of $S = 0, 1$ and $\infty$ for both the actual model and the bias correction and we obtain similar results.

Our conclusion is that, as long as a researcher is willing to say that the true $\rho$ stays between, say, .7 and 1.1, the delta prior is a better alternative than available bias corrections, even from a classical point of view.

6 Other Bayesian priors

Our paper contributes to the literature on priors for the VARs. We have discussed at length the advantages of the delta prior over the flat prior. Several alternatives to the flat prior in time series have been proposed and they usually also compensate the small sample bias. We start by commenting differences with the Minnesota, Jeffrey’s and shrinkage priors. More related to our work will be the discussion of Sims (1996) and Sims and Zha (1998) prior as well as the $S$-prior of Uhlig (1994).

6.1 Minnesota, Jeffrey’s and shrinkage priors

Applied work often uses the so-called Minnesota prior, with a Normal-Wishart prior centered at the unit root (Doan et al., 1984, see), which pushes the AR estimates towards a unit root. This estimator can be interpreted as delivering a bias correction, since in practice it usually increases the root of the process relative to OLS. But this correction is arbitrary, completely determined by the precision of the prior. It is difficult for an applied researcher to determine what are is his/her own beliefs about the parameters of the AR process and relate these to knowledge about how the economy works. This prior has been shown to be effective for forecasting purposes, but economists using a VAR to summarize the properties of the data generally avoid using this prior: if the empirical results are crucially determined by the precision of the prior, criticizing the prior becomes an obvious way to challenge the empirical results.

Phillips (1991) criticized the flat prior on the grounds that it was not truly uninformative. He suggested to use Jeffrey’s prior which, he argued, is truly uninformative and it appropriately gives larger weight to higher roots. By favoring higher roots, the prior implies some bias correction compared with OLS. This sparked an intensive debate, to which we have nothing to
add, but we will just note that Jeffrey’s prior is not commonly found in practical applications. In part, this is because of computational difficulties that make VAR applications hard. We just feel that most economists will feel more comfortable stating their prior about the growth rate than about parameters of the process. Also, our results show that criticism of the flat prior by Phillips (1991) on the grounds that the flat prior gave same weight to roots close to .5 or to 1 was missing the main point, since the problem are the huge initial growth rates. Our prior still gives the same weight to these alternatives but it does deliver a bias correction, and it reconciles the classical and Bayesian viewpoints.

Another alternative to the flat prior is the ‘shrinkage’ prior introduced for VARs by Ni and Sun (2003), which intends to curb overfitting by shrinking the parameters towards zero. As a result, conditionally on starting close to zero, it also constrains the initial growth rates implied by the parameters, so it ends up having a similar effect to our prior. Monte Carlo simulations of the authors suggest that on balance it tends to shrink mostly the constant terms, while somewhat pushing upwards the estimated slope coefficients, and therefore it corrects the small sample bias.

6.2 Priors similar to the delta prior

We have found two priors that happen to be close to the delta prior. Sims (1996) mentioned that, even though a flat prior justified using OLS from a Bayesian prospective, the results were unsatisfactory and OLS gave a systematic overestimation of the error in the first few observations. This is, in a way, observing the small sample bias and saying it should be corrected. Sims proposed to introduce a ‘dummy initial observation’ with no change in the variable. This is discussed in more detail in Sims and Zha (1998). This prior reflects the ‘belief that no-change forecasts should be ‘good’ at the beginning of the sample’ (Sims, 1996, p.5). It is, similarly as the Minnesota prior, an experience-based prior, introduced by adding a dummy observation with a variance. This prior is a special case of the delta prior when the prior mean is $g = 0$ and, somehow, the first $P$ observations are kept constant (perhaps by assuming that $\sigma_g = 0$ for $T_0 - 1$ periods and setting $\sigma_g = \sigma_\epsilon$ in the period $T_0$). Looking at the expression in our Result 1, it is clear that the delta posterior is consistent with a $\varphi$ that places probability 1 on such an evolution of $y$. In this sense, the prior of Sims (1996) is a special case of the delta prior. We feel that the delta prior is a more explicit way to express
the uncertainty about the data, that it has a clearer interpretation and it is easier for an applied researcher to relate to it.

Also close to the delta prior is the work of Uhlig (1994). It has been stated often that conditioning on the first few observations is incorrect, because, especially in a small sample, these initial observations carry valuable information about the parameters. Uhlig (1994) was an attempt at using the exact likelihood, putting the distribution of initial conditions in the likelihood.  

But, as should be clear from our discussion of modelling initial conditions above, there is not a clear way how to do this. Uhlig proposed to link parameters and initial conditions by saying how many periods the model has been operating (Uhlig’s S) and by stating what was the initial condition at the beginning of these periods. One alternative is to pick $S = \infty$ but this is as arbitrary as any other assumption: why not 10 periods?, why not since the end of WWII? If S is fixed, to what number? why assume the variable was at the centering parameter $\mu$ at the beginning? Furthermore, for roots equal or larger than 1 we have to assume a finite number of periods (usually one) and an arbitrary initial condition at that time (usually $y_{-1} = \mu$). It is probably because of this reason that the S-prior has not been used much in applied work.

In the appendix we report our Monte Carlo experiments with the S-prior in the univariate case, and we propose how to extend it to a VAR. It turns out, that the Uhlig’s S-prior has some of the nice properties of the delta-prior. In the appendix we show that introducing the S-prior delivers a similar asymmetry of the posterior, that $S = 0$ has very good properties in a classical sense, and that in an applied case with $S = 1$ it provides similar empirical results. We feel that these results reinforce the theme of the paper, namely, that if huge growth rates are discarded in the prior then a Bayesian observes an asymmetric posterior and much more precision can be gained from imposing such priors. It is because the S-prior also reduces the growth in the first few observations, and it makes transitions as in panel B of figure 1 very unlikely, that it has some features in common with the delta prior. However, we feel that the delta prior is easier to use for applied work.

16We call his approach a prior, but he stated it as writing the exact likelihood.
7 Christiano et al. (1999) estimation of the effect of monetary shocks in the US

To study the effect of bias and bias-correcting priors on macroeconomic VARs we first replicate the estimation of the effects of the monetary shock in the US from Christiano et al. (1999)\textsuperscript{17}, referred to further as CEE99. The authors estimate a VAR with output (Y, measured by the log of real GDP), prices (P, the log of the implicit GDP deflator), commodity prices (PCOM, the smoothed change in an index of sensitive commodity prices), federal funds rate (FF), total reserves (TR, in logs), nonborrowed reserves (NBR, in logs) and money (M1 or M2, in logs). All data are quarterly and the sample is 1965:3 - 1995:2. The residuals are orthogonalized with the Choleski decomposition of the variance (with this variable ordering) and the monetary shock is the one corresponding to the federal funds rate.

7.1 Impulse responses to monetary shocks

Figure 8 displays the responses of output to a monetary shock estimated with three procedures. The solid line is the responses from OLS estimator, equal across columns. The extreme dashed lines contain the 95% confidence bands and the middle line signals the median. The first column reproduces the results in CEE99 using a bootstrap procedure that is an attempt to provide an idea about uncertainty about the point estimates of impulse responses, while disregarding the small sample bias ('other-percentile' bootstrap). The second column is arguably the most popular way of constructing confidence bands, using the RATS procedure (Doan, 1992) for generating the Bayesian posterior bands for impulse responses assuming a flat prior. It is, therefore, centered at OLS and symmetric. The third column is the delta prior, conditional on the initial observations.

The bootstrap confidence bands for some series are displaced towards zero, compared with the point estimate, reflecting the small sample bias in time series. (The sample is quite large - 120 observations - so the bias may not be striking at first glance, but in smaller samples the bootstrap confidence bounds often exclude point estimates. Also, it is possible that this application has so many parameters that the sampling error is very large and the bias

\textsuperscript{17}We choose this example because the data is on the Internet, and because the authors use the bootstrap error bounds, which highlight the small sample bias.
is, by comparison, small potatoes. We are implementing these comparisons for some other applications where the bias may be larger.)

But in any case it is clear that the bias causes an underestimation of the effects of a monetary shock on output. The bias seems to be particularly large in the response of output and it is there also for the response of PCOM.

The next step is to impose the prior on initial growth rates of the variables. The parameters of the basic specification of the prior are given in table 1. We have chosen these values mechanically by setting the means equal simply to the average growth rates observed in the whole sample, and standard deviations are half of those observed. A more opinionated reader can plug in his/her own preferred values, but the results are not going to change much. Recall that in order to obtain the total variance one has to add the variance of the innovations, so the prior is quite loose. We set $T_0 = 10$.

<table>
<thead>
<tr>
<th>variable</th>
<th>prior mean annual growth rate</th>
<th>prior standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2.7</td>
<td>1.8</td>
</tr>
<tr>
<td>P</td>
<td>5.0</td>
<td>1.2</td>
</tr>
<tr>
<td>PCOM</td>
<td>3.2</td>
<td>100</td>
</tr>
<tr>
<td>FF</td>
<td>0.0</td>
<td>2.4</td>
</tr>
<tr>
<td>TR</td>
<td>5.4</td>
<td>4.5</td>
</tr>
<tr>
<td>NBR</td>
<td>5.2</td>
<td>3.3</td>
</tr>
<tr>
<td>M1</td>
<td>6.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>

In contrast to the non-informative prior, the delta-prior (see the plots in column 3 of figure 8) corrects the bias and typically shifts the posterior probability mass of the impulse responses away from zero. Notice that the 'other-percentile' bootstrap bands adjust in the wrong direction: in seeing the bootstrap in the upper left graph, the researcher should acknowledge that the bias reduces the absolute value of the response and, therefore, the researcher should report even more negative responses. But just displaying the bootstrap bands and median goes in the wrong direction. By contrast, the delta prior does the job automatically, and it gives more negative responses than OLS. The effect is strongest exactly where the bias, indicated by the deviation of the median from OLS, was strongest. Consequently, the
Figure 8: Impulse responses to monetary shocks: OLS point estimate, median and the 95% uncertainty bounds generated by 'other-percentile' bootstrap, RATS routine (flat prior), $\Delta$-prior (with parameters specified in the text) and $S = 1$-prior (the latter with trend, see the appendix)
responses of output estimated with the delta-prior are more negative and persistent, and the size of the interest rate shock is much more positive, giving more significance to the results of a VAR.

Another effect of the ∆-prior is that, tilting the estimates towards non-stationarity, it produces wider uncertainty bands than the non-informative prior, especially for longer lags. We believe that this reflects the fact that (despite the efforts of many researchers) it is virtually impossible, with a sample of 120, to ascertain the long run effect of a shock.\footnote{There is no paradox in the fact that the informative prior widens uncertainty bands, since their width depends both on the width and the location of the uncertainty bands of the autoregressive coefficients. Consider the response to a unit shock after 16 lags in the AR(1) model: $\rho \in (0.5, 0.7)$ corresponds to the response in the range $(0.000015, 0.0033)$, while a twice shorter range $\rho \in (1, 1.1)$ corresponds to responses in the range $(1, 4.6)$.}

Finally, we report what one would find with the $S=1$ prior for VAR’s, according to our interpretation of Uhlig’s prior in the Appendix. In this case, for the estimation to make sense, a trend must be included. The results are very close to the delta-prior. We feel this is reassuring since, as we said in section 6, the $S$-prior is another way to put very small probability on huge growth rates.

7.2 Comparison with Kilian (1998) bootstrap-after-bootstrap

Kilian (1998) proposed a classical procedure to construct error bands for impulse responses, called bootstrap-after-bootstrap. It is intended to improve the coverage of bootstrap intervals by approximately removing small-sample bias at each step of the simulation. Bootstrap is used to approximately remove the bias, and the underlying assumption is that the bias is constant in the neighborhood of the OLS estimate.

As described in Kilian (1998), the first step consists of estimating and correcting the bias of the OLS point estimate. In the second step, the error bands are generated: series are repeatedly simulated from the data generating process implied by the corrected OLS estimate, a VAR is estimated by OLS at each simulated data set, the OLS estimate is corrected for bias analogously as in the first step, and finally impulse responses are computed and stored. In a simplified, but computationally much cheaper version of this algorithm, the bias estimate obtained in the first step is reused in the second step, for correction of all OLS estimates on generated data (instead of performing a
Figure 9: CEE99 responses to monetary shocks: OLS estimates and percentiles 2.5, 50 and 97.5 of the distribution of impulse responses. Delta prior, 'Kilian' - bootstrap-after-bootstrap method and 'Kilian-nonst' - bootstrap-after-bootstrap method without correction of nonstationarity.
separate bootstrap for each of them). We use this simplified bootstrap-after-bootstrap here.

Bias correction pushes the roots of the process towards the nonstationary range and explosive responses may often result. This may, however, be a spurious effect caused by the assumption of constant bias. Kilian suggests to shrink bias estimates in all cases, when the bias corrected estimate would become explosive, and thus guarantee stationarity of all the simulated distribution of impulse responses. Sims and Zha (1998), when applying Kilian’s method, deviate here and allow for explosiveness. We report results from both approaches.

The first column in 9, labeled 'delta prior' shows again, for easier comparison, results obtained with delta prior. The second column, labeled 'Kilian', shows results obtained with Kilian’s simplified bootstrap-after-bootstrap method, constraining roots of bias-corrected OLS estimates to be stationary. The results are quite similar to the ‘naive bootstrap’ bands from figure 8: the error bands are narrow and in some cases (output and federal funds rate) displaced towards zero, displaying the small sample bias towards stationarity. The third column, labeled 'Kilian-nonst' (for 'nonstationary'), shows the error bands computed with simplified bootstrap-after-bootstrap, but without imposing stationarity. The results are more similar to the results of the delta-prior estimation, where stationarity is not imposed either. The resulting bands reflect it, becoming very wide for higher lags. Overall, the figure illustrates the dilemma involved in applying bootstrap-after-bootstrap, when the root of the system is close to unity: imposing stationarity discards much of the bias correction. Allowing nonstationarity, on the other hand, exposes the results to the inaccuracy of the assumption of constant bias. We guess that it is because of the failure of this assumption in practice, that the bands contain too much nonstationarity, and are so wide for farther lags. It is possible that applying the full, not simplified, version of Kilian’s algorithm gives better results.

8 Conclusions

One theme of the paper has been that as long as huge growth rates are given small weight there is no dichotomy between Bayesian and classical estimation in time series. Both with the $S$-prior and our delta prior the posterior looks asymmetric and the same issues that had been discussed in
short sample classical analysis arise in Bayesian econometrics. A Bayesian adjusts upwards (towards nonstationarity) the OLS estimator in a similar way as corrections proposed in the classical literature.

Our preferred alternative is the delta prior. It should be uncontroversial that most economic variables can not have huge growth in any given period; introducing this knowledge about the economy reconciles a Bayesian posterior with the certainty that classical econometricians (at least those that correctly worry about short sample issues) have about the fact that OLS should be adjusted upwards close to a unit root.

But correcting the bias turns out not to be a very good way to estimate parameters and, at the same time, take care of short sample issues. It seems that a Bayesian estimator based on a delta prior is also the best from the classical point of view.

We have illustrated the effect of the delta-prior in a practical case: the estimation of the macroeconomic effects of a monetary policy shock, following Christiano et al. (1999). The posterior corrects the small sample bias towards stationarity, and implies stronger responses, in particular of output and interest rates, than those found by the authors.

This version of the paper is very much just a progress report. We have to implement the exact posterior calculations and explore further implications of the delta prior. In particular, it will be interesting to examine the forecasting performance of the delta-prior compared with the Minnesota prior and other standard forecasting tools. Second, it will be interesting to see the effect of delta-prior in other structural VAR applications, where short sample issues may be important.

References


**APPENDIX I: Prior from the initial conditions: the $S$-prior**

**A $S$-priors for the AR(1)**

In this section we explore the implications of the initial conditions, formulated following Uhlig (1994). Uhlig considers processes which had started at the mean $S$ periods before the first observation in the sample.\(^\text{19}\)

\(^{19}\)Strictly speaking, Uhlig (94) does not derive a 'prior' from the above specification of the initial condition, but he uses it to write the 'exact likelihood'. This is, of course, a semantic issue, since the resulting posterior is the same as if the above specification for the initial condition is considered a prior given the initial observations, which is our preferred interpretation.
Let us first consider the case $S = 0$ and call a fourth helicopter to tour a joint p.d.f. Now, the initial condition become

$$y_0 = \mu$$ \hspace{1cm} (24)

with probability one, which is analogous to the condition $y_0 = 0$ of Sims and Uhlig (1991).

Two views of the resulting joint distribution of $\rho, \hat{\rho}$ are displayed in figures 10 and 11. Figure 10 shows the fixed-$\rho$ cross-section, which is the small sample distribution of $\hat{\rho}$. It is skewed towards lower values of $\hat{\rho}$, because of the small sample bias. Comparing with Figure 2 we can see that this bias is stronger when the constant term is estimated. We can also see that, as $\rho$ increases, the distribution becomes tighter, but not as fast as that in Figure 2.

Figure 11 shows the fixed-$\hat{\rho}$ cross-section which is the Bayesian posterior of $\rho$ given $\hat{\rho}$. Consistently with the observation of Andrews (1993), the balance between the downward bias of OLS, and its increasing precision, is tilted towards the former. The posterior accounts for this and is skewed towards values of $\rho$ higher then the observed $\hat{\rho}$, so that, in this figure, when we condition on an observed value $\hat{\rho} = 1$, the mean of the posterior is larger than one. Therefore, in this case we also have that there is no strong dichotomy between Bayesian and classical analysis: both recommend correcting for the bias of OLS.

The general 'S-prior' (Uhlig (1994) and (Zivot, 1994, p.569-570)) is:

$$p(\mu|\sigma, y_0, S) = \sigma^{-1} \nu(\rho, S)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \nu(\rho, S)^{-1} (y_0 - \mu)^2 \right\}$$ \hspace{1cm} (25)

with

$$\nu(\rho, S) = \frac{(1 - \rho^2S)}{(1 - \rho)} \quad |\rho| \neq 1$$

$$= S \quad |\rho| = 1$$ \hspace{1cm} (26)

These priors are then complemented with the non-informative prior for $\rho$ and $\sigma$:

$$p(\rho, \sigma) \propto \sigma^{-1}$$ \hspace{1cm} (27)

The $S = 1$ prior is conditionally conjugate, and the full posterior distribution can be conveniently simulated with the Gibbs sampler. For the general $S$ the conjugacy is lost, but the marginal posterior for $\rho$ can be obtained
Figure 10: Joint frequency distribution of $(\rho, \hat{\rho})$ sliced along the $\rho = 1$ line (as in Figure 3 from Sims and Uhlig (1991)). Model with a constant term, starting from $y_0 = \mu$

Figure 11: Joint frequency distribution of $(\rho, \hat{\rho})$ sliced along the $\hat{\rho} = 1$ line (as in Figure 4 from Sims and Uhlig (1991)). Model with a constant term, starting from $y_0 = \mu$
analytically by a straightforward integration of the posterior with respect to \( \mu \) and \( \sigma \). The resulting formula is given in Zivot (1994, equation (43)).

The bias and mean squared error of the Bayesian ‘\( S \)-prior’ estimators are examined in a Monte Carlo experiment similar to the one in section 5. We denote the Bayesian posterior means obtained with \( S = 0 \), \( 1 \) and \( 100 \) priors respectively as BS0, BS1 and BS100.

We simulate the AR(1) processes with initial conditions corresponding to \( S = 0 \), \( 1 \) and \( \infty / 1 \) and in each case try all 3 Bayesian estimators, in addition to OLS and the CBC. The results are presented in figures 12, 13 and 14.

The BS1 and BS100 estimators perform somewhat better, but are broadly similar to the delta prior: in RMSE terms they beat OLS for the positive \( \rho \) and the CBC on an even larger interval (except that the BS100 has slightly worse RMSE than the CBC for \( \rho \) around 1). In the range where they perform well, their bias (unreported here) is stronger than that of the CBC, although reduced compared with OLS. They are not very sensitive to the accuracy of the prior, e.g. the BS1 estimator performs similarly also when the prior assumption is false: for the processes starting at the mean (Figure 12), and

![Figure 12: RMSE of the OLS, BS0, BS1 and BS100 estimators in a Monte Carlo experiment with \( S = 0 \), sample size: \( T = 25 \).](image)
Figure 13: RMSE of the OLS, BS0, BS1 and BS100 estimators in a Monte Carlo experiment with $S = 1$, sample size: $T = 25$.

Figure 14: RMSE of the OLS, BS0, BS1 and BS100 estimators in a Monte Carlo experiment with $S = \infty$, sample size: $T = 25$. 
in the stationary distribution (Figure 14).

That last observation is not true for the BS0 estimator, which gives a particularly strong advantage over the alternatives when the prior is correct (for the processes starting exactly at the mean, Figure 12), but for some parameter values it is very misleading when the prior is wrong. The graphs are truncated, so it cannot be seen that the RMSE of BS0 has a peak of 0.65 (more than 3 times that of the alternatives) around \( \rho = -0.6 \) for \( S = 1 \) (Figure 13) and of 0.9 (more than 6 times that of the alternatives) around \( \rho = -0.9 \) for \( S = \infty / 1 \) (Figure 14). Nevertheless, in the range which is most relevant in practice, i.e. for high positive values of \( \rho \), BS0 is always more precise than the alternatives.

B The \( S \)-prior for VAR’s

In this section we generalize the ‘\( S=1 \)’ prior to the VAR(P) model, evolving around an arbitrary exogenous process (for example a constant mean, or a linear trend).

Let \( Y \) be a \( T \times J \) matrix gathering \( T \) observations on \( J \) jointly endogenous variables. Let \( L \) denote a lag operator and define a \( P \times 1 \) vector \( l \equiv [L, L^2, \ldots, L^P]' \). Define \( X \equiv l' \otimes Y = [Y_{-1}, Y_{-2}, \ldots, Y_{-P}] \), a \( T \times JP \) matrix with \( P \) lagged matrices \( Y \). \( W \) is a \( T \times K \) matrix with \( K \) exogenous variables (in case of a time invariant mean, \( W = \iota_T \), a \( T \times 1 \) vector of ones, while in case of a linear trend \( W = [\iota_T, \tau] \), where \( \tau = [1, 2, \ldots, T]' \). \( Z \equiv l' \otimes W = [W_{-1}, W_{-2}, \ldots, W_{-P}] \) is a \( T \times KP \) matrix with \( P \) lagged matrices \( W \). Finally, \( U \) is a \( T \times J \) matrix with \( T \) independent normal vectors of length \( J \), each with a variance \( \Sigma \).

Deviations of \( Y \) from the exogenous component evolve according to a VAR(P) process, perturbed by shocks \( U \):

\[
Y - W\Gamma = (X - Z (I_P \otimes \Gamma)) \Pi + U
\]

where \( \Pi \) is a \( JP \times J \) matrix of VAR coefficients and \( \Gamma \) is a \( K \times J \) matrix with means and slopes of the time trend for each endogenous variable. Reduced form of the above model is:

\[
Y = W\tilde{\Gamma} + X\Pi + U
\]

with

\[
\tilde{\Gamma} = \Gamma - (l' \otimes \Gamma)\Pi = \Gamma (I - \Phi_1L - \ldots - \Phi_P L^P)'
\]

43
where $\Phi_i$ is a matrix of VAR coefficients of lag $i$, so that $\Pi = (\Phi_1, \Phi_2, \ldots, \Phi_P)'$. Equation (30) is a generalization of the restriction $\alpha = \mu(1 - \rho)$ (20) for the AR(1) case, and analogously to that case it has implications when the lag polynomial in $\Phi$s contains some unit roots. When $\Gamma$ consists only of constant terms, (30) guarantees that those corresponding to series that have unit roots are suppressed, and the random walks have no drifts. When $\Gamma$ contains constant terms and time trends, (30) ensures that the time trends corresponding to unit root series are suppressed, consistently with the postulate of Uhlig (1994).

Likelihood conditional on initial $P$ observations is proportional to:

$$L(Y; \Pi, \Gamma) \propto |\Sigma|^{-T_2} \exp \left\{ -\frac{1}{2} \text{tr} \Sigma^{-1}U'U \right\}$$  (31)

A common noninformative prior for the error variance and the autoregressive coefficients is:

$$p(\Sigma, \Pi) \propto |\Sigma|^{-\frac{1}{2}(J+1)}$$  (32)

The prior for $\Gamma$ must be proper, to compensate the indeterminacy at unit root implied by restriction (30). A 'bias-correcting' prior assumption, in the spirit of the previous sections, relates the coefficients of the deterministic component of the process to the initial observations:

$$Y_0 - W_0\Gamma = U_0$$  (33)

where $Y_0$ is a $T_0 \times J$ matrix with $T_0$ initial observations on the endogenous variables, $W_0$ is a $T_0 \times K$ matrix with the corresponding values of the exogenous variables, and $U_0$ a $T_0 \times J$ matrix with $T_0$ independent normal vectors length $J$ with a variance $\Sigma_0$. We take $\Sigma_0$ to be the OLS estimate of the VAR errors.

The implied prior for $\Gamma$ is normal, centered on the OLS estimate of $\Gamma$ on the $T_0$ initial observations:

$$p(\text{vec}\Gamma|Y_0, W_0, \Sigma_0) = N(\text{vec}\hat{\Gamma}_0, \Sigma_0 \otimes W_0'W_0^{-1})$$  (34)

where

$$\hat{\Gamma}_0 \equiv (W_0'W_0)^{-1}W_0'Y_0$$

The simplifying assumption in (33), which delivers conjugacy of the prior, is that the first $T_0$ deviations from the deterministic component are independent, and only starting with time period 0 the time dependence implied by
Π kicks in. An exact multivariate and multilag counterpart to the Uhlig’s S-prior would assume, that the process had been equal to its deterministic component up to period $-S$ and then started evolving according to (28). But then the variance of $U_0$ would depend in a complicated way on Π and the conjugacy of the prior would be lost. The prior in (34) implies that the initial $T_0$ observations are used only to infer on the coefficients of the deterministic process Γ, but not the Π. This may be a simplification, but it is still better then the case of the flat prior, where the information in the initial observations is ignored altogether.

The joint posterior implied by (31),(32) and (34) is:

$$p(Π, Γ, Σ|Y, Y_0) ∝ |Σ|^{-\frac{T+J+1}{2}} \exp\left\{-\frac{1}{2} tr(Σ^{-1}U'U + Σ_0^{-1}U'_0U_0)\right\}$$

(35)

The conditional posteriors are:

$$p(Σ|Π, Γ, Y, Y_0) = IW(U'U, T)$$

(36)

$$p(vecΠ|Γ, Σ, Y, Y_0) = N(vec̃Π, Σ ⊗ (̃X'̃X)^{-1})$$

(37)

$$p(vecΓ|Π, Σ, Y, Y_0) = N(G^{-1}g, G^{-1})$$

(38)

where IW is an Inverted Wishart distribution,

$\tilde{X} ≡ X - Z(I_P ⊗ Γ)$

$\tilde{Π} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'(Y - WΓ)$

$G = A'(Σ^{-1} ⊗ I_T)A + Σ_0^{-1} ⊗ W'_0W_0$

$g = A'(Σ^{-1} ⊗ I_T)vec(Y - XΠ) + (Σ_0^{-1} ⊗ W'_0W_0) vec\hat{Γ}_0$

$A = I_J ⊗ W - (Π' ⊗ Z)((I_P ⊗ K_{JP})(vecI_P ⊗ I_J) ⊗ I_K)$

and $K_{JP}$ is the commutation matrix (Magnus and Neudecker, 1988, p.46). A sample from the posterior (35) can be generated by Gibbs sampler, i.e. by drawing in turn from (36), (37) and (38).

APPENDIX II: Some derivations

C Integrating constant for a multivariate normal likelihood

For completeness we put here the derivation of the integrating constant given in (11). Throughout we condition on a known error variance Σ.
Likelihood conditional on initial $P$ observations is equal to:

$$L(Y; B) = \left(2\pi\right)^{-\frac{N}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left(-\frac{1}{2} \operatorname{vec}(Y - XB)'(\Sigma^{-1} \otimes I_T) \operatorname{vec}(Y - XB)\right) =$$

$$(2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left(-\frac{1}{2} \left(\operatorname{vec} B'(\Sigma^{-1} \otimes X'X) \operatorname{vec} B - 2 \operatorname{vec} B' \operatorname{vec}(X'Y\Sigma^{-1}) + \operatorname{tr} Y'Y\Sigma^{-1}\right)\right)$$

(39)

It is easy to see that this expression is proportional to the multivariate normal density of $B$: $N\left(\operatorname{vec}(X'X)^{-1}X'Y, \Sigma \otimes (X'X)^{-1}\right)$ or

$$p(B|Y) = \left(2\pi\right)^{-KJ/2} |\Sigma|^{-K/2} |X'X|^{J/2}$$

$$\exp\left(-\frac{1}{2} \operatorname{vec}(B - (X'X)^{-1}X'Y)'(\Sigma^{-1} \otimes X'X) \operatorname{vec}(B - (X'X)^{-1}X'Y)\right) =$$

$$(2\pi)^{-KJ/2} |\Sigma|^{-K/2} |X'X|^{J/2}$$

$$\exp\left(-\frac{1}{2} \left(\operatorname{vec} B'(\Sigma^{-1} \otimes X'X) \operatorname{vec} B - 2 \operatorname{vec} B' \operatorname{vec}(X'Y\Sigma^{-1}) + \operatorname{tr} Y'X(X'X)^{-1}X'Y\Sigma^{-1}\right)\right)$$

(40)

By definition, the normalizing constant is such that:

$$p(B|Y) = \frac{L(Y; B)}{K_f(Y)}$$

and therefore the formula for the normalizing constant is:

$$K_f(Y) = \int L(Y; B)dB = \frac{L(Y; B)}{p(B|Y)} =$$

$$(2\pi)^{-\frac{(T-K)J}{2}} |\Sigma|^{-\frac{T-K}{2}} |X'X|^{-\frac{J}{2}} \exp\left(-\frac{1}{2} \operatorname{tr} Y'(I - X(X'X)^{-1}X')Y\Sigma^{-1}\right)$$

$$= (2\pi)^{-\frac{(T-K)J}{2}} |\Sigma|^{-\frac{T-K}{2}} |X'X|^{-\frac{J}{2}} \exp\left(-\frac{1}{2} \operatorname{tr} (\hat{E}'\hat{E}\Sigma^{-1})\right)$$

(41)