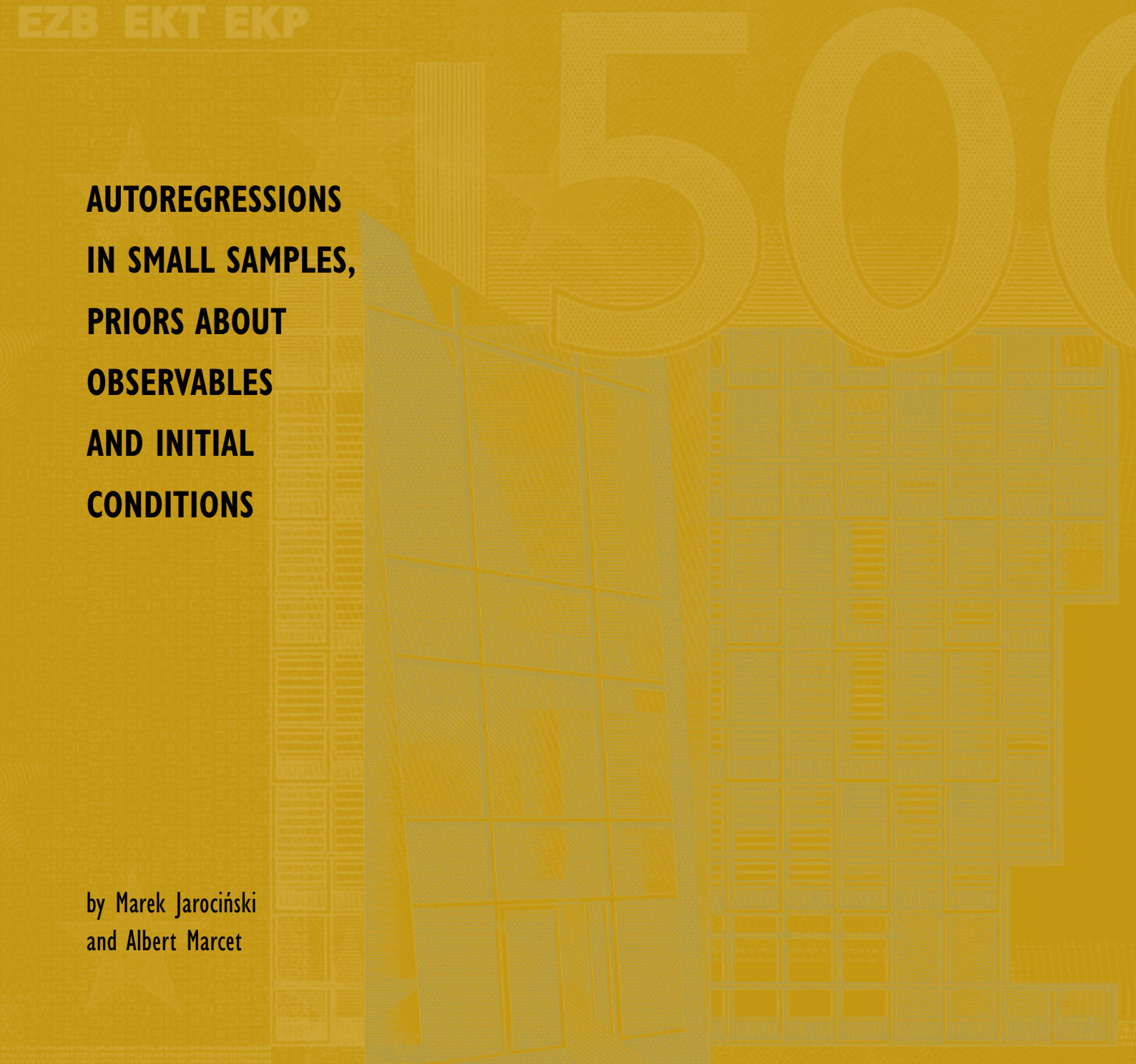
EUROPEAN CENTRAL BANK
EUROSYSTEM

WORKING PAPER SERIES
NO 1263 / NOVEMBER 2010



AUTOREGRESSIONS IN SMALL SAMPLES, PRIORS ABOUT OBSERVABLES AND INITIAL CONDITIONS

by Marek Jarociński
and Albert Marcet

EUROPEAN CENTRAL BANK

EUROSYSTEM

WORKING PAPER SERIES

NO 1263 / NOVEMBER 2010

AUTOREGRESSIONS IN SMALL SAMPLES, PRIORS ABOUT OBSERVABLES AND INITIAL CONDITIONS[1]

by Marek Jarociński[2]

and Albert Marcet[3]

In 2010 all ECB publications feature a motif taken from the €500 banknote.

This paper can be downloaded without charge from http://www.ecb.europa.eu or from the Social Science Research Network electronic library at http://ssrn.com/abstract_id=1699149.

1 We thank Gianni Amisano, Manolo Arellano, Stéphane Bonhomme, Matteo Ciccarelli, Jean Pierre Florens, Oliver Linton, Bartosz Maćkowiak, Peter C.B. Phillips, Thomas J. Sargent, Frank Schorfheide and Chris Sims for useful conversations, and Harald Uhlig for his comments on the earlier version of this paper. All errors are our own. Albert Marcet acknowledges financial support from CREI, DGES (Ministerio de Educación y Ciencia), CIRIT (Generalitat de Catalunya) and European Community FP7-SSH grant MONFISPOL under grant agreement SSH-CT-2009 225149. All computations were performed using Ox. Part of the work was included in a previous working paper of ours titled "Prior on Growth Rates, Small Sample Bias and the Effects of Monetary Policy".

2 European Central Bank, Kaiserstrasse 29, D-60311 Frankfurt am Main, Germany; email:marek.jarocinski@ecb.europa.eu

3 London School of Economics, CEP and CEPR; e-mail: a.marcet@lse.ac.uk

Address
Kaiserstrasse 29
60311 Frankfurt am Main, Germany

Postal address
Postfach 16 03 19
60066 Frankfurt am Main, Germany

Telephone
+49 69 1344 0

Internet
http://www.ecb.europa.eu

Fax
+49 69 1344 6000

Information on all of the papers published in the ECB Working Paper Series can be found on the ECB's website, http://www.ecb.europa.eu/pub/scientific/wps/date/html/index.en.html

ISSN 1725-2806 (online)

CONTENTS

Abstract

We propose a benchmark prior for the estimation of vector autoregressions: a prior about initial growth rates of the modeled series. We first show that the Bayesian vs frequentist small sample bias controversy is driven by different default initial conditions. These initial conditions are usually arbitrary and our prior serves to replace them in an intuitive way. To implement this prior we develop a technique for translating priors about observables into priors about parameters. We find that our prior makes a big difference for the estimated persistence of output responses to monetary policy shocks in the United States.

Keywords: Vector Autoregression, Initial Condition, Bayesian Estimation, Prior about Growth Rate, Monetary Policy Shocks, Small Sample Distribution, Bias Correction

JEL codes: C11, C22, C32

Non-technical summary

We address a long-standing problem in econometrics: how to estimate autoregressive models with few observations. Univariate and vector autoregressive (VAR) models are commonly used in applied research. However, the existing econometric methods either have well known deficiencies, or rely on arbitrary and/or non-transparent assumptions. This paper proposes a new approach which incorporates available information in a transparent way and combines the advantages of Bayesian and of classical estimators.

The most common approach to the estimation of autoregressions is to use the ordinary least squares (OLS) estimator. However, OLS tends to underestimate persistence in autoregressive models when a small sample is available. This may significantly affect empirical results, especially the impulse responses at longer lags. For a classical econometrician the problem with OLS is manifested in its *bias*, which is known since 1950s. This bias carries with it a large *mean squared error* of the OLS estimator. In parallel, the Bayesian literature has also identified a number of problems with using the posterior distribution centered at the OLS estimate.

Classical and Bayesian econometricians have produced a plethora of techniques to estimate autoregressions in small samples with the aim to correct the problems that OLS is known to have. However, it is safe to say that there is no widely accepted alternative in practice. As a consequence many applied papers still use OLS in highly overparameterized vector autoregressions (VARs). Any deviation from OLS, whether inspired by classical or Bayesian procedures, is liable to criticism for having made certain ad hoc choices that may be (and often are) crucial for the results. The aim of this paper is to design a widely acceptable procedure for estimating autoregressions with small samples.

This paper begins by re-examining (in section 2) the following well known puzzle: despite the small sample bias in autoregressions, OLS is the best estimator for a Bayesian with a flat prior and a standard quadratic loss function. This poses a disturbing dilemma to an applied economist: Classical econometrics tells him that the true process is more persistent than his OLS estimate. In contrast, Bayesian econometrics seems to tell him to use the OLS estimate, unless he has an informative prior. In many standard applications classical and Bayesian-flat-prior inference are identical, in spite of philosophical differences. But the OLS in autoregressions is an example where the two schools, Bayesian and classical, deliver contradicting conclusions. So should an "objective" researcher correct the OLS estimate upwards or not? One interpretation is that this differing view between Bayesian-flat-prior and a classical econometrician is ok, since they belong to different camps. This is a disappointing conclusion for most applied economists who do not have strong views about the fundamentals of statistics.

This paper delivers a comforting message: We first show that the recommendations of the Bayesian and classical econometrics are not as far apart as commonly believed. It turns out that the contradiction resulted from the fact that by default classical analysis uses very different assumptions on initial conditions from those used in the default Bayesian analysis. In fact, given the same treatment of the initial condition, Bayesian and classical econometricians roughly agree about the appropriateness or not of OLS. And, in particular, both of them would adjust the OLS estimate towards non-stationarity for "reasonable" treatments of the initial observations.

Therefore, it is fundamental to relate the initial observations to parameters. Again, there is a myriad of alternatives for modelling initial conditions in the literature. We propose instead to relate initial observations and parameters in a truly Bayesian way and to incorporate information in our estimates that most reasonable economists do have. Our proposal is to use an informative prior based on the a priori distribution about the observed series in the first few periods of the sample. For example, if GDP is one of the variables in a VAR, the analyst should ask the following question to his/her client "what is your a priori distribution for GDP growth?". Most economists will definitely have a well formed opinion about likely values of GDP growth. The answer to this question can and should be incorporated in the posterior distribution. To a classical econometrician our approach is simply a strategy for specifying initial conditions of the process. To a Bayesian this means specifying a Bayesian prior given the observed initial condition.

The key advantage of our approach is that it requires stating a prior about observables. This makes it relatively easy to express the answer, as economists often have a good intuition about the behaviour of economic variables. Disagreements among economists about this may be small and we can reach a near-consensus prior. In contrast, the standard Bayesian approach is to specify priors about parameters. Parameters in VAR models are not easily interpretable and therefore it is extremely difficult to specify intuitive priors about them.

However, a substantial technical difficulty arises in using a priori information about observable time series. The prior on observables has to be "translated" into a prior on coefficients, which is the one used in Bayes' rule. This is an operation that involves solving an integral equation. The existing techniques for solving such equations are not appropriate in our case due to the very high dimension of the parameter space in VARs and because we are only interested in approximate solutions. In section 3 we design a numerical algorithm based on a fixed point formulation of the integral equation. We show that this algorithm works very well in various empirical applications.

Section 4 discusses two empirical applications. First, we show how our prior affects the estimates of persistence of stock prices from the well known Extended Nelson-Plosser dataset. This application serves to show that the alternatives available in the literature give many different results with little guidance for choosing among them, and it shows how our estimate compares with others. The second example is the famous study by Christiano, Eichenbaum and Evans (1999) on the effects of monetary policy shocks on output in the US. This serves to show that even in a large-scale VAR the algorithm we propose works and it makes a significant difference. The effect of monetary shocks on output is much higher when estimated with our approach, our estimate of the effect on output doubles the one reported by Christiano, Eichenbaum and Evans (1999). We conclude that monetary shocks are much more important than had been previously thought.

In section 5 we show that our estimator has also desirable properties from a classical econometrics point of view. We study the frequentist performance of an estimator constructed as the posterior mean obtained with our approach. We show that this estimator performs better than other classical bias corrected estimators in important dimensions, even when judged by classical criteria.

Therefore, applied economists may find our approach very attractive, regardless of their views about the validity of Bayesian or classical approaches.

1 Introduction

The ordinary least squares (OLS) estimator tends to underestimate persistence in autoregressive models when a small sample is available. This may significantly affect empirical results, especially the impulse responses at longer lags. For a frequentist the problem with OLS is manifested in its bias and a large mean squared error, known since 1950s.[1] Bayesians also tend to be dissatisfied with the flat prior posterior, centered at the OLS estimate.[2] Many techniques have been designed to estimate autoregressions in small samples using both classical and Bayesian approaches. However, it is safe to say that there is no widely accepted way to proceed. In fact, many applied papers still use OLS in highly overparameterized vector autoregressions (VARs). Any deviation from OLS, whether inspired by classical or Bayesian procedures, is liable to criticism for having made certain ad hoc choices that may be (and often are) crucial for the results.

Our aim is to design a widely acceptable procedure for estimating autoregressions with small samples. We begin by reexamining the following well known puzzle: despite the small sample bias in autoregressions, OLS is the best estimator for a Bayesian with a flat prior and quadratic loss.[3] Hence, a classical econometrician concerned with small sample issues and a Bayesian have very different views about the validity of OLS in autoregressions. One interpretation is that there is no puzzle: Bayesians and frequentists belong to different camps, differences in estimation are natural. We find this is a disappointing conclusion, as many applied economists do not have strong view about Bayesian vs classical approach.

In section 2 we show that there is no such puzzle: in fact, *given the same treatment of the initial condition,* Bayesian and classical econometricians agree about the appropriateness or not of OLS. And, in particular, both of them would adjust the OLS estimate towards non-stationarity for "reasonable" treatments of the initial observations.

Therefore, it is fundamental to relate the initial observations to parame-

[1]The earliest references are Quenouille (1949), Hurwicz (1950), Marriott and Pope (1954) and Kendall (1954). A general characterization of the effects of the bias on the highest root is in Stine and Shaman (1989). Abadir et al. (1999) show that the bias becomes more severe in multivariate models.

[2]See e.g. Phillips (1991), Uhlig (1994b), Sims and Zha (1998), Sims (2000). Sims (2000) argues that under a flat prior posterior the initial condition explains an unreasonably large share of the variation. Sims and Zha (1998, p.959) refer to the excessive stationarity of the flat-prior posterior as "the other side of the well-known bias toward stationarity of least-squares estimates of dynamic autoregressions."

[3]This has been known for a long time. Sims and Uhlig (1991) revived this point and illustrated it with graphical and analytic arguments.

ters. Again, there is a myriad of alternatives for modeling initial conditions in the literature. We propose to relate initial observations and parameters in a truly Bayesian way and to incorporate information that most economists do have. Our proposal is to use an informative prior based on the a priori distribution of the observed series in the first few periods of the sample. For example, if GDP is one of the variables in a VAR, the analyst should ask the following question to his/her client "what is your a priori distribution of GDP growth in the beginning of the sample?". The answer to this question should be incorporated in the posterior distribution.

This kind of prior has many advantages: i) it allows to clearly relate initial observations and parameters, as required by our previous discussion, ii) it may be a near consensus prior: a room full of economists is sure to be full of disagreements, but the range of opinions about the prior distribution of GDP growth is bound to be relatively narrow, and whatever differences remain will have a clear interpretation, iii) it is much easier to express an opinion about a prior distribution of observed variables than of VAR parameters, iv) it entirely sidesteps the issue of what is a "truly" uninformative prior in time series,[4] we prefer to use priors that are indeed informative but that are widely acceptable.

The usefulness of thinking about priors on observables is brought out by reexamining the validity of the flat prior from this, purely Bayesian, perspective. A flat prior about the parameters in a VAR with a constant term corresponds to an a priori belief that the growth rate of GDP in the first few periods is very likely to exceed, say, 100%! Researchers routinely use flat priors in the hope that such priors are neutral and yield posteriors close to posteriors from reasonable subjective priors. But in this paper we show many examples where posteriors with reasonable subjective priors differ significantly from the flat prior posteriors. Therefore, estimating a VAR by OLS is unjustified unless one genuinely has crazy beliefs about initial growth rates.

Another great advantage of our prior approach is that it should be highly appealing to a frequentist. In section 5 we study the frequentist performance of an estimator constructed as the posterior mean obtained with a purely data-driven prior about initial growth rate. We show that from a purely frequentist point of view this estimator is an attractive alternative to other classical bias corrected estimators. Therefore, applied economists who do not have strong views about the validity of Bayesian or classical approaches should be at ease with our approach.

A substantial technical difficulty arises in using a priori information about

[4]This issue has been raised by Phillips (1991).

observable time series, because the standard Bayesian analysis requires a prior about parameters, not about observables. A classical discussion of priors specified in terms of observables can be found in Berger (1985, Ch.3.5). The prior about observables has to be "translated" into a prior about coefficients. This operation involves solving a Fredholm integral equation. There has been a recent interest in the microeconometrics literature in solving these *inverse problems*.[5] The techniques used in this literature, as well as those discussed by Berger, can not be applied directly to our case due to the very high dimension of the parameter space in VARs and because we are only interested in approximate solutions to the relevant Fredholm equation. In section 3 we design an algorithm based on a fixed point formulation of the problem. We show that this algorithm works very well in various empirical applications.

Section 4 discusses two empirical applications. First, we show how our prior affects the estimates of persistence of stock prices from the well known Extended Nelson-Plosser dataset. This application shows that the alternatives available in the literature give many different results with little guidance about which to choose, and it shows how our estimate compares with others. The second example is the famous study by Christiano et al. (1999) on the macroeconomic effects of monetary policy shocks. This example shows that even in a large-scale VAR the algorithm we propose works and it influences the results in a significant way: the effect of monetary shocks on output is much higher using our approach than using that of Christiano et al. (1999).

A very large literature is concerned with the issues tackled in our paper. The frequentist literature has proposed many methods for correcting the OLS estimator in small samples.[6] But each correction focuses only on some aspects of the problem (for example, it focuses on the bias of a specific transformation of the parameters). Construction of confidence intervals for these estimators is tricky.[7] Decision-theoretical justifications of these approaches are questioned.[8]

Applied Bayesians dissatisfied with the flat prior have used priors which are supposed to push the posterior towards unit roots, such as the famous

[5] See Carrasco et al. (2007) for a summary of such applications.

[6] Some examples of such estimators are Quenouille (1949), Orcutt and Winokur (1969), Andrews (1993), MacKinnon and Smith (1998), Kilian (1998) and Roy and Fuller (2001). A large literature using local to unity asymptotics is also justified in terms of it small sample properties.

[7] See e.g. Mikusheva (2007) and references therein.

[8] Berger and Wolpert (1988) discuss how a concern about frequentist properties of statistical procedures can lead to unreasonable inferences.

Minnesota prior or dummy observations priors.[9] However, these priors are rarely seen as actually representing prior knowledge and they are often considered ad hoc.[10] Furthermore, the Minnesota prior can give rise to paradoxical behavior, sometimes pushing the posterior away from the unit root, as it does in our example in section 4.1. One of the contributions of our approach is that it provides a rationale for dummy observation priors, as we show that they are equivalent to priors about growth rates with particular variances.

The importance of the initial condition in estimating autoregressions with small samples has also been discussed before.[11] What is new in our paper is the point that the treatment of the initial condition is what drives much of the disagreement about OLS between frequentists and flat-prior Bayesians. The literature on the so-called "exact likelihood" is one attempt to relate parameters and the initial observation. But this approach has well recognized problems that we discuss in section 2.4. It rests on many ad hoc assumptions and it is rarely used in applied work. Instead we focus on informative priors about the initial behavior of the series which, we think, are much more likely to generate consensus and which achieve the same goal of relating initial condition and parameters.

Priors stated in terms of observables are rare in the time series literature. Kadane et al. (1996) use priors about one period ahead forecasts. Villani (2009) uses a prior about the unconditional long run mean of growth rates and his prior can be specified directly for parameters in a reparameterized VAR. His VAR is in differenced variables and not in levels, which implies a dogmatic prior about the low frequency behavior.

The paper is organized as follows. In section 2 we discuss the role of the initial condition in classical and Bayesian estimation of autoregressions. This is useful to motivate our prior about initial growth rates. In section 3 we discuss translating priors about observables into a prior distribution of model parameters. In section 4 we present two empirical applications. Finally, in section 5 we present a frequentist evaluation of our prior in the case of the

[9] See Doan et al. (1984), Uhlig (1994b), Sims and Zha (1998) and Sims (2006).

[10] Sims (2000, p.452) recognizes that these priors are unsatisfactory, when he concludes: "There are open research questions here, and few well-tested procedures known to work well in a wide variety of applications. More research is needed - but on how to formulate reasonable reference priors for these models, not on how to construct asymptotic theory for nested sequences of hypothesis tests that seem to allow us to avoid modeling uncertainty about low-frequency components."

[11] The role of the initial condition is discussed among others in Blundell and Bond (1998); Chamberlain (2000); Arellano (2003) from the classical perspective and in Schotman and Van Dijk (1991); Uhlig (1994a); Sims (2000) from the Bayesian perspective. DeJong et al. (1992) and Müller and Elliott (2003) show how the initial condition determines the power of frequentist unit root tests.

AR(1) process. We conclude in section 6.

2 OLS: Classical vs Bayesian or Initial Condition?

We first argue that the differing views between classical and Bayesian-flat-prior approach about whether to correct OLS are driven entirely by a different treatment of the initial condition. Throughout this section we use as example an AR(1) model with an intercept:

$$y_t = \alpha + \rho y_{t-1} + u_t, \quad t = 1 \dots T \tag{1}$$

where u_t is i.i.d. $N(0, \sigma_u^2)$. OLS estimates are denoted $(\alpha^{OLS}, \rho^{OLS})$.

2.1 A Puzzle

The following facts about the adequacy of ρ^{OLS} are well known:

- *Frequentist view*: for given values of (α, ρ) and with ρ near 1, the small sample distribution of ρ^{OLS} is skewed to the left and its mean is lower than ρ. An example of this density is displayed with the solid line in Figure 1 for $\rho = .95$ and $T = 100$.
- *Bayesian view*: under a flat prior for (α, ρ) the posterior distribution of ρ is symmetric and centered, precisely, at ρ^{OLS}. This posterior is represented by the dashed line in Figure 2.[12]

These facts imply that a classical econometrician proceeds very differently from a Bayesian econometrician who has a flat prior. Classical econometricians concerned about small sample performance have designed various corrections for OLS. On the other hand, many empirical papers using OLS to estimate autoregressions with small samples justify this estimator by invoking the flat prior. This contrast is intriguing and it is disturbing for practitioners who do not have a strong preference towards either frequentist or Bayesian approach with weak priors.[13]

[12] This dashed line is in fact an average of all posteriors that correspond to $\rho^{OLS} = 0.95$. It is obtained with an approximately flat prior, see further discussion.

[13] While it is controversial what priors are really weak in the appropriate sense (see Phillips, 1991, and comments in the same issue of the Journal of Applied Econometrics), the flat prior remains the baseline and it is justified as a tool for reporting the shape of the likelihood.

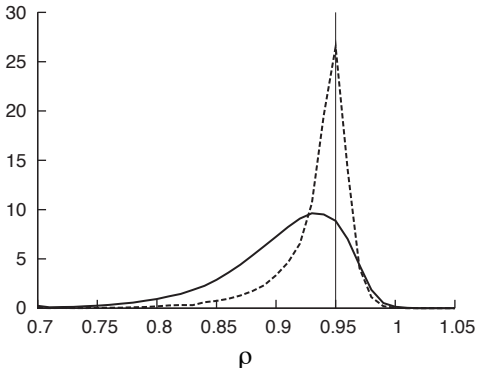


Figure 1 – **Frequentist Density.** Density of ρ^{OLS} conditional on $\rho = 0.95$. Initial condition (2) with $\sigma_0^2 = \sigma_u^2 \sum_{i=0}^{S-1} \rho^{2i}$ (continuous line) and $\sigma_0^2 = 100\sigma_u^2 \sum_{i=0}^{S-1} \rho^{2i}$ (dashed line). We take $S = 100$. Sample size is $T = 100$. Construction of these densities is explained in Appendix A.

2.2 Initial Condition, Just a Detail?

To make the discussion concrete, assume the following relation between the initial observation and parameters:

$$y_0 = \alpha \left(\sum_{i=0}^{S-1} \rho^i \right) + u_0 \quad \text{with} \quad u_0 \sim \mathrm{N}(0, \sigma_0^2) \tag{2}$$

with S and σ_0^2 given. This condition can be justified by assuming that the process started at time $-S$ with a known initial value $y_{-S} = 0$. The first term of (2) is the deterministic component of the process at time 0 and u_0 is the stochastic component. Throughout this section we assume finite S. A large variety of alternatives for modeling initial conditions are available in the literature, we review them in section 2.4. Some papers assume instead $S = \infty$, reparameterize the constant term α as $\mu(1-\rho)$ or/and use separate assumptions when $\rho \geq 1$. Our main results are not affected by these details.

To a frequentist equation (2) gives the distribution of the initial observation given parameter values for α, ρ. Most frequentist studies consider a variance σ_0^2 that is related to the parameter values considered. One common assumption is that the shocks in periods $\{-S+1, \ldots, 0\}$ have the same distribution as the shocks in periods $\{1, \ldots, T\}$, i.e. $N(0, \sigma_u^2)$ so that the variance of the stochastic component $\sigma_0^2 = \sigma_u^2 \sum_{i=0}^{S-1} \rho^{2i}$. We have used this value of σ_0^2 to construct the density represented by the solid line in Figure 1.

To a Bayesian (2) specifies a restriction on the prior about α, ρ given the observed y_0. For example, assume a flat prior for ρ, $p(\rho) \propto 1$. Then (2) can be used to derive the prior density $p(\alpha|\rho, y_0, \sigma_0^2)$. It is clear that the flat prior

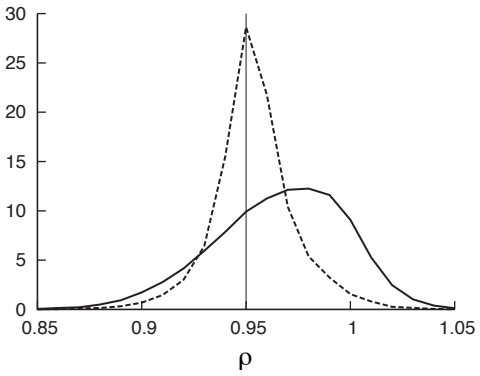


Figure 2 – **Bayesian Density.** Densities of ρ conditional on $\rho^{OLS} = 0.95$. Initial condition (2) with $\sigma_0^2 = \sigma_u^2 \sum_{i=0}^{S-1} \rho^{2i}$ (continuous line) and $\sigma_0^2 = 100\sigma_u^2 \sum_{i=0}^{S-1} \rho^{2i}$ (dashed line). We take $S = 100$. Sample size is $T = 100$. Construction of these densities is explained in Appendix A.

corresponds to taking $\sigma_0^2 = \infty$. Correspondingly, the dashed line in Figure 2 shows the posterior density of ρ conditional on an estimate ρ^{OLS} for a very large σ_0^2 (specifically, 100 times the σ_0^2 used in the previous paragraph). As expected for a nearly flat prior, the posterior is nearly symmetric and centered at ρ^{OLS}.[14]

This highlights that a flat-prior Bayesian analysis differs from the standard frequentist small sample approach not only with the Bayesian treatment of data and parameters, but also in the treatment of the initial condition. Flat-prior Bayesian takes extremely large σ_0^2 while frequentists tend to use "reasonable" values for σ_0^2. Which of these differences is responsible for the puzzle mentioned above?

2.3 The role of the initial condition

To see the impact of the initial condition let us now reverse the assumptions about σ_0^2 between the frequentist and the Bayesian.

2.3.1 Frequentist analysis with large σ_0^2

Consider now the frequentist distribution of ρ^{OLS} when the stochastic component of the initial condition has $\sigma_0^2 = \infty$, as in the flat-prior initial condition. It is known, but rarely highlighted, that in this case the small sample bias

[14] We choose the size of σ_0^2 for better graphical representation, to ensure that both densities are still clearly visible on the same graph. As σ_0^2 increases, the density becomes more symmetric, but also more peaked.

vanishes.[15] This can be seen in Figure 1. The dashed line shows the frequentist distribution of ρ^{OLS} when $\rho = .95$ but the initial condition is drawn from (2) with the same very large value of σ_0^2 that we used in the case of the near-flat prior shown in Figure 2. It is clear that the bias becomes much smaller and the small sample distribution of the OLS estimator becomes concentrated near the true value.

Using such a large σ_0^2 is probably unreasonable, but this discussion serves to show that both a frequentist and a flat-prior-Bayesian agree that OLS is a good estimator as long as they model initial conditions with a large σ_0^2. The reason why OLS is a good estimator for large σ_0^2 is illustrated in Figure 3. Each row of graphs represents a realization of $y_1, ..., y_T$, with the same sequence of shocks $u_1, \ldots, u_T$ in both rows, but different realized u_0 in each row: in the top row $u_0 = 0$ while in the bottom row u_0 is large and negative. The left column of graphs plots y_t against time. The process is stationary and the transition from the remote starting value to the steady state dominates the dynamics of the series in the lower row.

The right column shows a scatterplot of the right-hand-side variable (y_{t-1}) against the left-hand-side variable (y_t) in the regression on equation (1) for each sample. This is the "cloud of points" that undergraduate econometrics books display to show how a regression line fits the data. The solid line in the scatterplots is the regression line implied by the true parameters in (1) while the dashed line is the fitted regression implied by the parameters estimated by OLS for this realization. The slope of the dashed line is lower than the actual regression line. The lower slope in the top-right graph reflects the OLS bias which usually results in $\rho^{OLS} < \rho$ for the present parameter values. The slope in the bottom-right graph, however, almost coincides with the true regression line reflecting the result mentioned in the first paragraph of this subsection.

These graphs make it clear why the initial condition is the key: the explanatory variable (y_{t-1}) shows much higher dispersion in the bottom row of graphs. Realizations like the one in the bottom row are more common when the density of the initial observation is more spread out, ie. when σ_0^2 is large. In such realizations OLS is a good estimator even for a frequentist, because the sample variance of the explanatory variable y_{t-1} is large. As is well known, a high variance of the explanatory variable means that OLS is a good estimator. This is why the fitted regression in the bottom row is much closer to the true regression line.

[15] This result can be found in Phillips (1987, section 6) and Phillips and Magdalinos (2009). Also Arellano (2003, p.86) and Chamberlain (2000) point this result for some special cases in the context of panel data.

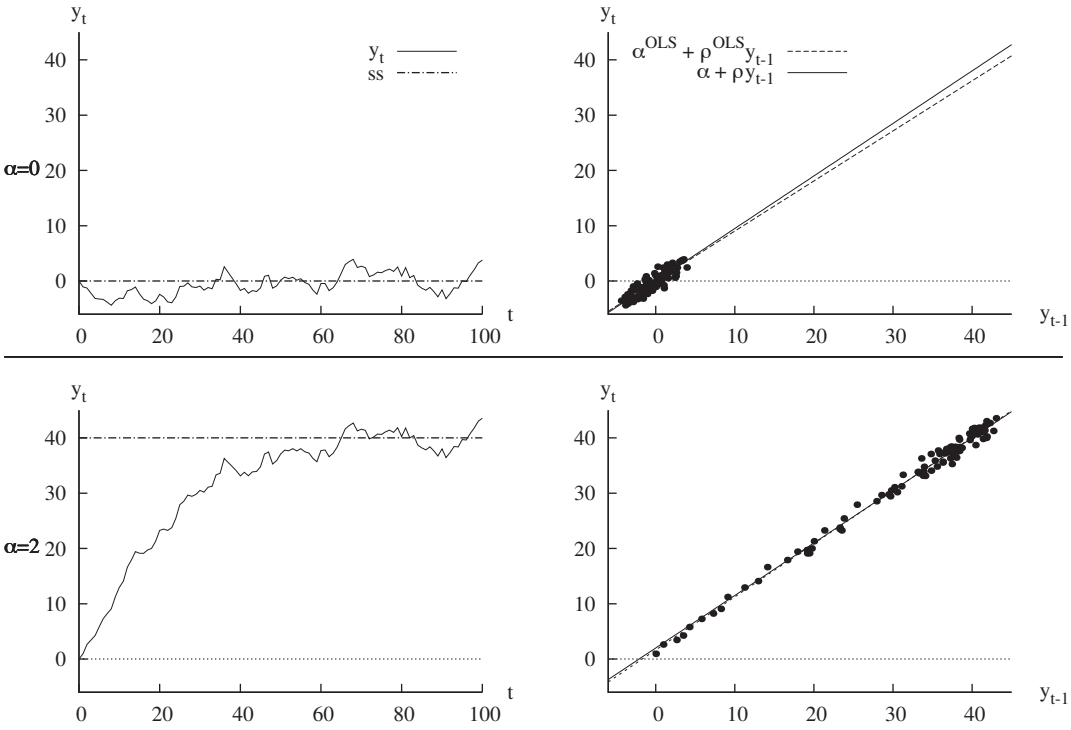


Figure 3 – Two cases of the AR(1) process and the performance of the OLS estimator of the coefficients. The left column plots y_t against time. The right column shows scatter plots of y_t against y_{t-1}, along with true and fitted regression lines.

2.3.2 Bayesian analysis with small σ_0^2

We now incorporate a small σ_0^2, standard in the frequentist literature, into a Bayesian analysis. This is also the approach of Bayesian papers using the so called "exact likelihood"[16] and papers that specify a prior for the parameters conditional on the initial observation.[17]

Figure 2 illustrates the effect of σ_0^2 on the posterior beliefs about ρ. The continuous line represents the density of ρ given an observed ρ^{OLS} when we take $\sigma_0^2 = \sigma_u^2 \sum_{i=0}^{S-1} \rho^{2i}$. This Bayesian density is clearly asymmetric and its mean is higher than the OLS estimate. In this example, upon observing $\rho^{OLS} = 0.95$ a Bayesian would believe that the true parameter ρ is around 0.97, adjusting the OLS estimate upwards, in the same direction as a frequentist concerned about the bias. This illustrates that when a reasonable initial condition is assumed, Bayesians tend to agree with the frequentists

[16] See e.g. Zellner (1971, ch.7.1), Uhlig (1994a) and Lubrano (1995).
[17] See e.g. Schotman and Van Dijk (1991).

that the OLS estimate is too low and they also correct it upwards.

While the argument given here is purely numerical and for a specific example, we give an analytic result for a related case in the next subsection, see our discussion after Result 1.

To our minds this resolves the puzzle we mentioned in section 2.1: classical and Bayesian econometricians qualitatively agree about the virtues of OLS near unit roots when they model the initial observation analogously. For large σ_0^2 they agree that OLS is great. But for small σ_0^2 they agree that ρ^{OLS} should be adjusted upwards.

2.4 The delta prior

The above discussion suggests that it is crucial to have a plausible joint distribution relating the initial observation y_0 and the parameters α, ρ when we only have small samples.

One possibility would be to find a "good" model of the initial condition, that is, to find the best possible specification of equations such as (2), and apply the so-called "exact likelihood" approach. The literature has proposed very many alternative ways of specifying this initial condition. These alternatives differ in the number of periods S for which the model was holding in the past, in the way that past uncertainty enters, in what to do for parameter values $\rho \geq 1$, etc. To cite a few: the initial condition we consider in section 2.2 is one of the cases of Uhlig (1994a). Andrews (1993) takes $S = \infty$ for $\rho \in (-1, 1)$ but he uses an arbitrary initial condition at $\rho = 1$ and rules out $\rho > 1$. Bhargava (1986) and MacKinnon and Smith (1998) assume $S = \infty$ in the deterministic component when $|\rho| < 1$ and they assume $\alpha = 0$ when $\rho = 1$; for the stochastic component they assume $S = 1$, i.e. they take $\sigma_0^2 = \sigma_u^2$. Phillips and Magdalinos (2009), on the other hand, assume $\sigma_0^2 = \kappa\sigma_u^2$ for a certain κ.

It is obviously very difficult to choose from among these options. For example, the most widely used approach is to take $S \to \infty$ for $|\rho| < 1$, but this amounts to making the identifying assumption that the model and its parameters have been stable for infinitely many periods before the start of the sample. This assumption is often implausible in practice. Moreover, this approach gives rise to a disturbing discontinuity at $\rho = 1$. An alternative would be to agree on a reasonable value S for which the model has been stable, but it would be difficult to build consensus on a reasonable value for S. Then one needs another identifying assumption about y_{-S}. Then one still needs to choose one of the myriad alternative specifications of the stochastic component.

There is no guidance to choose among these alternatives. We find this

approach unsatisfactory and given how few papers have used the initial condition in applications of VARs we are probably not alone.

Our proposal is to use a purely Bayesian approach and specify the prior beliefs about the behavior of the series for the first few periods. Since we condition on the initial observation y_0, this prior relates the initial observation to parameters as is required by the discussion in Section 2.3. In the case of non-stationary variables it is most natural to state a prior about growth rates, so we will concentrate on this prior in the discussion, but it would be trivial to adapt our arguments for the use of prior information about the likely *level* of a variable instead.

This approach has several advantages. First, it seems much easier to build a (near-) consensus view about reasonable values of growth rates for many variables than about the initial condition used in the "exact likelihood". Another advantage is that this prior is easy to elicit: it should be easy for most economists to express their views about the likely behavior the growth rate of, say, GDP. Also, such conditional growth rate is well defined regardless of the model being stationary or not, so the discontinuity at a unit root is entirely avoided. Most economists do have strong opinions about likely behavior of many variables, ignoring this knowledge amounts to throwing away relevant information in VARs that are very often highly parameterized.

Let us say that we ask an economist about his/her prior beliefs regarding the growth rate of some series in period $t = 1$. Letting y represent the log of the series, the answer might be expressed as

$$\Delta y_1 \sim \mathrm{N}(\mu_\Delta, \sigma_\Delta^2) \tag{3}$$

for some values $\mu_\Delta, \sigma_\Delta^2$. For reasons to be discussed below, we need to assume $\sigma_\Delta^2 > \sigma_u^2$.[18] This economist should think about what he/she thought about this series back in period $t = 0$, when the sample started, given knowledge of y_0. We leave conditioning on y_0 implicit in the above notation.

The prior information (3) should be incorporated in the posterior of α, ρ. One difficulty in forming this posterior is that (3) is not a prior statement about the distribution of parameters, as required in Bayes' rule. For this purpose we have to translate (3) into a prior distribution of parameters α, ρ. To clarify our semantics: throughout the paper we call a statement such as (3) a "prior about observables" and we reserve the name "delta prior" for the implied distribution of unobservable parameters.

Prior (3) implies

$$(1 - \rho)y_0 - \alpha = u_0 \tag{4}$$

[18] The assumption of normality in the prior is convenient in this section. However, the techniques discussed in section 3 and applied in section 4 do not need normality.

for $u_0 \sim \mathrm{N}(\mu_\Delta, \sigma_\Delta^2 - \sigma_u^2)$. Hence, this is a restriction on the joint distribution of y_0 and the parameters derived from the prior knowledge about the analyzed series.

For simplicity, in the rest of this section we assume $y_0 = 0$. In the AR(1) case translating the prior (3) is very easy and it implies the following prior about the constant term

$$\alpha \sim \mathrm{N}(\mu_\Delta, \sigma_\Delta^2 - \sigma_u^2). \tag{5}$$

The following result characterizes the effect of this prior on the posterior mean of ρ. A complete formula for this mean is given in equation (B.2) in the Appendix. Let $Y^T \equiv [y_1, ..., y_T]$.

Result 1. *Assume a flat prior* $p(\rho) \propto 1$ *and any prior* $p(\sigma_u^2)$ *defined for* $\sigma_u^2 > 0$.

If $\rho^{OLS} < 1$ *and* $\left(\frac{y_T - y_0}{T} - \mu_\Delta\right)$ *is sufficiently small, then*

$$E(\rho|Y^T) > \rho^{OLS}. \tag{6}$$

Result 1 implies that the Bayesian with a delta prior thinks that the persistence of the process is higher than the OLS estimate. That is, he/she adjusts the OLS estimator in the same direction as frequentists worried about the small sample bias. The conditions needed for this result are very weak: $\rho^{OLS} < 1$ includes the range where the frequentists are most concerned with the bias (many papers in the frequentist literature design estimators explicitly only for the range $\rho^{OLS} < 1$ but close to 1, see e.g. Roy and Fuller (2001)). The condition that the sample mean growth rate $\frac{y_T - y_0}{T}$ is close to the prior mean growth rate μ_Δ simply means that the prior is "reasonable", not too different from the observed growth rate.

Result 1 is perhaps surprising, since the delta prior only concerns α for the case considered, when $y_0 = 0$.

Result 1 is illustrated in Figure 4, showing the Bayesian density of $\rho|\rho^{OLS}$ for two different delta priors. The figure conditions on a realization $\rho^{OLS} = 0.95$. The dashed line shows the density of $\rho|\rho^{OLS}$ when the prior growth rate has mean zero ($\mu_\Delta = 0$) and standard deviation 6.5%, while the dotted line takes a growth rate of 3% ($\mu_\Delta = 0.03$). As Result 1 suggests, we can see that both densities shift the mean to the right: for $\mu_\Delta = 0$ we find $E(\rho|\rho^{OLS}) = 0.97$ and for $\mu_\Delta = 0.03$ we find $E(\rho|\rho^{OLS}) = 0.952$, both higher than $\rho^{OLS} = 0.95$.

This Figure is also useful to highlight the great advantage of using priors about growth rates: a discussion about which is the relevant posterior distribution in Figure 4 turns into a discussion about what is a reasonable value

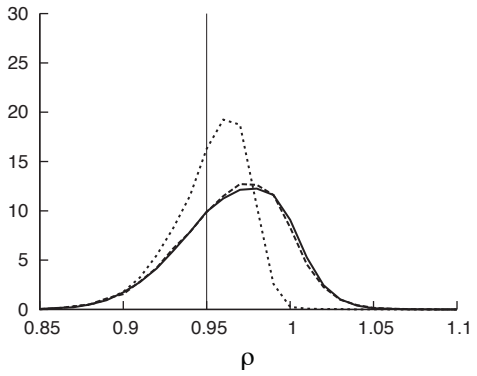


Figure 4 – **Bayesian posteriors.** Densities of ρ conditional on $\rho^{OLS} = 0.95$. Initial condition (2) for $\sigma_0^2 = \sigma_u^2 \sum_{i=0}^{S-1} \rho^{2i}$ with $S = 100$ (continuous line); delta prior with $\mu_\Delta = 0, \sigma_\Delta = 0.065$ (dashed line); delta prior with $\mu_\Delta = 0.03, \sigma_\Delta = 0.065$ (dotted line). In all cases $\sigma_u = 0.057$ and $T = 100$. Construction of the these densities is explained Appendix A.

a priori for $E_0(\Delta y_1)$. The answer, of course, depends on the exact series at hand. For example, if y represents log real GNP most economists are likely to disagree with $\mu_\Delta = 0$, they might prefer to state that GNP is likely to grow, for example, at 3% and therefore prefer the dotted line in Figure 4. This matters for inference since, as we can see from the figure, the posterior with $\mu_\Delta = 0.03$ is concentrated on values closer to the OLS estimate and the implied upward correction of ρ^{OLS} is smaller.

This also allows us to compare the delta prior with previously used Bayesian approaches. For example, the exact likelihood approach in a Bayesian framework (as in section 2.3.2) implies that, a priori, the researcher thinks that $E_0(\Delta y_1) = 0$. We would probably think that this is not appropriate for GNP, and we would prefer the delta prior with $\mu_\Delta = 0.03$ instead. We can also discuss the appropriateness of the widely used flat prior. This amounts to assuming $\sigma_0^2 = \infty$ in (2) and it implies a prior belief that the growth rate is very likely to be larger than, say, 100%. Since this is a prior belief that no analyst would ever hold about GNP and since it matters for estimation (because only in this case OLS is justified) we hope that applied economists will never again use OLS in autoregressions including GNP.

Result 1 also gives analytic support to our claim in section 2.3.2 that the posterior mean under exact likelihood is higher than OLS. So far we only backed this claim with a numerical result shown in Figure 2, which made a standard assumption on the initial condition. It is easy to check that the delta posterior when $\mu_\Delta = 0$ is equal to the exact likelihood posterior under the less standard assumption on the initial condition that σ_0^2 is given

and independent of ρ (for example, this is the assumption in Phillips and Magdalinos (2009)) and $S = 0$. Although the initial condition we used in Figure 2 is slightly different, we can see that the continuous line of Figure 4 which repeats the continuous line of Figure 2, is actually very similar to that with the delta prior and $\mu_\Delta = 0$.

In this section it was easy to translate the prior analytically because the prior about growth rates (4) only involved one period $t = 1$. But in models with many parameters stating a prior for only one period amounts to throwing away a lot of information and it is desirable to use priors about more periods. In this case analytic solutions such as (5) are not available, which motivates the next section.

3 Translating Priors

We discuss how to translate a general prior about *observables* into a prior distribution of *unobservable* parameters. The numerical method that we propose can be used to incorporate other information on VARs, for example arising from experience or formal economic models, or it can be used in other time series models.

3.1 Defining the prior for parameters

Let us consider a general N-dimensional stochastic process $\{y_t\}$. We define $Y^T \equiv [y_1, ..., y_T]'$ as a $T \times N$ matrix gathering the random variables from which the sample of T observations is drawn. A model (say, a VAR with a given lag length) determines the likelihood function known to the researcher, $p_{Y^T|B}(\overline{Y}^T; \overline{B})$ ($\overline{Y}^T$ and $\overline{B}$ denote realizations of random variables Y^T and B). The observed initial observation is a parameter entering the functional form of $p_{Y^T|B}$.

We assume that the researcher is willing to state a prior density about $Y \equiv [y_1, ..., y_{T_0}]'$, the studied variables in periods $1, \ldots, T_0$ for some T_0 which needs not be equal to T. (For consistency, we should be using Y^{T_0} but we omit the superscript for brevity). This density will be denoted as p_Y. It represents what the researcher thinks before observing the sample about the likely behavior of the series in the first T_0 periods. It is, therefore, a marginal density of the observable data in the first T_0 periods. The likelihood function of Y, consistent with the same model as before, will be denoted as $p_{Y|B}$. The uncertainty represented in p_Y is a combination of the researcher's uncertainty about the actual values of parameters B and the error terms of the model in $p_{Y|B}$.

Let $\mathcal{B}$ be the space of possible parameters B and let $\mathcal{Y}$ be the space of possible values for Y. It is clear that knowledge of $p_{Y|B}$ and p_Y places the following restriction on the marginal density of the parameters p_B:

$$\int_{\mathcal{B}} p_{Y|B}(\overline{Y};\cdot)\, p_B = p_Y(\overline{Y}) \quad \text{for almost all } \overline{Y} \in \mathcal{Y} \tag{7}$$

This equation says that the joint density of observables Y and parameters B, integrated over the parameters, has to equal the marginal density of Y as specified by the prior p_Y. Our task will be, given the known density p_Y and the likelihood $p_{Y|B}$, to find the prior density p_B that satisfies the functional equation (7). Equations of this type are known in calculus as Fredholm equations of the first kind and in statistics as inverse problems.

The above problem may not have any solution for some pairs of p_Y and $p_{Y|B}$. For example, we already pointed out in the AR(1) case analyzed in section 2.4 that in order for (3) to be consistent with the model (1) we needed a prior variance $\sigma_\Delta^2 \geq \sigma_u^2$. If this is violated, the researcher's belief in p_Y is incompatible with the model $p_{Y|B}$, the researcher is asking the model to do something it can not do. As we will see later, in practice this needs not be a problem because even if the exact solution does not exist, one may be able to find a prior for parameters which approximately delivers the desired distribution p_Y to a satisfactory degree.

Another possibility is that the above problem has multiple solutions. This is likely to be the case e.g. when the dimension of B is larger than the dimension of Y, as in section 2.4 above. In this case equation (7) delivers only a restriction on the prior. Then the researcher needs to complete the prior with a density of the so far unrestricted dimensions, which can be e.g. flat if no additional prior knowledge is available. Therefore, multiplicity of solutions of (7) needs not be a problem.

The Bayesian econometrics literature has paid little attention to the issue of translating priors. An exception is Chapter 3.5 of Berger (1985).[19] The techniques used in this literature and those used to solve Fredholm equations are usually designed to obtain very accurate solutions to relatively low-dimensional problems. They often involve solving non-linear systems of equations with gradient methods that would be unfeasible for our purpose since standard VARs involve very high-dimensional problems.[20] Furthermore, we

[19] A related approach is the 'predictive approach' to elicitation, where a prior about observables takes the form of a statement about one step ahead predictive density conditional on the known right hand side variables (Kadane et al., 1980). This approach has been applied in the time series context in Kadane et al. (1996). Translating the prior in this literature was typically an easy step.

[20] For example, the prior distribution of the coefficients in the Christiano, Eichenbaum and Evans (1999) model that we discuss in section 4.2 has more than 20,000 parameters.

are not too worried about matching p_Y exactly. In practice, researchers would rarely have very strong views about the exact mean, variance and functional form of the prior about initial growth rates of the modeled series, and they are probably willing to accept Bayesian analysis with a slightly different prior. Therefore it is enough to find a prior for parameters that matches "reasonably well" the specified prior about growth rates. What is "reasonable" is also subjective, but much easier to specify when we talk about prior mean of output growth than, say, the mean of an autoregressive parameter on the fourth lag of the variable.

3.2 Fixed point formulation

We now reformulate the problem of translating the prior from observables to parameters as the solution to a fixed point problem. This formulation suggests a practical algorithm to find an approximate prior by successive iterations.

Let $g : \mathcal{B} \to \mathcal{R}_+$ be any density on B. Define the functional $\mathcal{F}_{p_Y}$ as

$$\mathcal{F}_{p_Y}(g)(\overline{B}) \equiv \int_{\mathcal{Y}} \frac{p_{Y|B}(\overline{Y};\overline{B})\ g(\overline{B})}{\int_{\mathcal{B}} p_{Y|B}(\overline{Y};\cdot)\ g}\ p_Y(\overline{Y})\ d\overline{Y} \quad \text{for all } \overline{B} \in \mathcal{B} \tag{8}$$

Clearly $\mathcal{F}_{p_Y}(g) : \mathcal{B} \to \mathcal{R}_+$ is itself a density, hence $\mathcal{F}_{p_Y}$ maps the space of densities for B into itself. $\mathcal{F}_{p_Y}(g)$ has the following interpretation: the term

$$p^g_{B|Y}(\overline{B}|\overline{Y}) \equiv \frac{p_{Y|B}(\overline{Y};\overline{B})\ g(\overline{B})}{\int_{\mathcal{B}} p_{Y|B}(\overline{Y};\cdot)\ g}$$

is the posterior obtained if the prior on parameters is g and if the data realization $\overline{Y}$ would be observed. Therefore, $\mathcal{F}_{p_Y}(g)$ is a mixture of posteriors for different realizations $\overline{Y}$, each weighted by its probability $p_Y(\overline{Y})$.

Applying this functional repeatedly is like learning better and better about the parameters B by repeatedly computing posteriors given samples drawn from p_Y. In the fixed point of such iteration the parameters prior p_B is fully consistent with the observables prior p_Y and with the model. The relationship between this fixed point and problem (7) is given in the following proposition:

Proposition 1. *If p_B satisfies (7), then p_B is a fixed point of $\mathcal{F}_{p_Y}$.*

Proof

We show if p_B solves (7) then $\mathcal{F}_{p_Y}(p_B) = p_B$. We have for all $\overline{B} \in \mathcal{B}$

$$\mathcal{F}_{p_Y}(p_B)(\overline{B}) = \int_{\mathcal{Y}} p_{Y|B}(\overline{Y};\overline{B})\ p_B(\overline{B})\ d\overline{Y} = p_B(\overline{B}) \int p_{Y|B}(\cdot;\overline{B}) = p_B(\overline{B})$$

The first equality holds from the definition of $\mathcal{F}$ and (7), the second equality takes $p_B(\overline{B})$ before the integral since it does not depend on $\overline{Y}$. The last equality holds because $p_{Y|B}$ is a density so it integrates to 1 over $\mathcal{Y}$.□

3.3 Gaussian approximate fixed point

The previous proposition suggests that p_B might be found by iterating on $\mathcal{F}_{p_Y}$ in order to find a fixed point. This means iterating on densities. Only in very special cases these iterations can be performed analytically: we discuss one such special case in Appendix D. We propose here a numerical method that has worked for us in all practical applications we have tried.

For the case of a normal likelihood we start at a normal density for B and iterate on $\mathcal{F}_{p_Y}$ only approximately, always staying within the realm of normal densities along the iterations. This is convenient for three reasons. First, it guarantees that along the way we always have a proper density for B.[21] Second, a normal density is fully described by its mean and variance, which greatly reduces the dimensionality of the problem. Third, and most important, for normal error terms u and normal priors we have closed form formulas for the posteriors $p^g_{B|Y}$ involved in the definition of $\mathcal{F}_{p_Y}$, and we use these closed form formulas to speed up the calculations. The prior p_Y can take any form as long as Monte-Carlo draws from it can be easily computed.

We now make some assumptions to build the likelihood. We assume in the rest of the paper that $\{y_t\}_{t=1}^T$ is a VAR(P) process:

$$y_t = \sum_{i=1}^{P} \Phi_i \; y_{t-i} + \gamma + u_t \qquad t = 1, ..., T \tag{9}$$

$u_t \sim \mathrm{N}(0, \Sigma_u)$ i.i.d., Φ_i are $N \times N$ matrices and γ is vector of N constant terms (generalizing this to the case with other exogenous variables is straightforward). The VAR(P) for $t = 1, ..., T_0$ can be written in matrix form as

$$Y = XB + U \tag{10}$$

Here Y is defined as at the beginning of section 3, X collects lagged values of Y in the usual way and it also has a column of ones which multiplies the

[21] A well-known problem with trying to solve Fredholm equations is that common approximation schemes fail because the approximation may fall outside the admissible set of functions. For example, discretizing $p_{Y|B}$, p_B and p_Y and solving (7) as a system of linear equations is known to fail because it is likely to yield an approximate discrete p_B that is not a distribution and because it is very sensitive to small changes in the discretization scheme.

constant terms, $B \equiv [\Phi_1, \ldots, \Phi_P, \gamma]'$ and $U \equiv [u_1, ..., u_{T_0}]'$. We assume for simplicity the error variance Σ_u to be known.

Throughout this section we always condition on P actual initial observations $\mathbf{Y}^0 \equiv [\mathbf{y}_{-P+1}, ..., \mathbf{y}_0]'$ but we omit the symbol "$|\mathbf{Y}^0$" for brevity. Then $p_{Y|B}$ is the standard conditional likelihood function of a gaussian VAR:

$$p_{Y|B} = N_{\mathrm{vec}\, Y}\left((I_N \otimes X)\,\mathrm{vec}\, B, (\Sigma_u \otimes I_{T_0})\right) \tag{11}$$

Note that $\mathbf{Y}^0$ is contained in the first N rows of X.

We now look for successive approximate iterations on the mapping $\mathcal{F}_{p_Y}$ within the space of normal distributions. A well known result in Bayesian econometrics is that given a gaussian prior $g = N(\mu_g, \Sigma_g)$, with some mean μ_g and variance Σ_g, the posterior $p^g_{B|Y}$ conditional on observing a value $\overline{Y}$ is also gaussian with variance and mean

$$\mathrm{Var}_{p^g(\cdot|\overline{Y})}(B) = \left(\Sigma_g^{-1} + \Sigma_u^{-1} \otimes \overline{X}'\overline{X}\right)^{-1} \tag{12}$$

$$E_{p^g(\cdot|\overline{Y})}(B) = \mathrm{Var}_{p^g(\cdot|\overline{Y})}(B)\ \left(\Sigma_g^{-1}\mu_g + \mathrm{vec}(\overline{X}'\overline{Y}\Sigma_u^{-1})\right) \tag{13}$$

This implies that $\mathcal{F}_{p_Y}(g)$ is a mixture of normal distributions $p^g_{B|Y}$, which is not in general a normal distribution. However, we approximate $\mathcal{F}_{p_Y}(g)$ itself with a gaussian distribution with the mean and variance of B under the distribution $\mathcal{F}_{p_Y}(g)$. This mean and variance can be found with a Monte Carlo procedure based on the following result:

Result 2. *Given any g, for any function $h : \mathcal{B} \to R^m$ we have*

$$E_{\mathcal{F}_{p_Y}(g)}(h(B)) = E_{p_Y}\left[E_{p^g(\cdot|Y)}(h(B))\right] \tag{14}$$

and, in particular,

$$E_{\mathcal{F}_{p_Y}(g)}(B) = E_{p_Y}\left[E_{p^g(\cdot|Y)}(B)\right] \tag{15}$$

$$Var_{\mathcal{F}_{p_Y}(g)}(B) = E_{p_Y}\left(Var_{p^g(\cdot|Y)}(B)\right) + Var_{p_Y}\left[E_{p^g(\cdot|Y)}(B)\right] \tag{16}$$

Proof [22]

$$E_{\mathcal{F}(g)}(h(B)) = \int_{\mathcal{B}} h(\overline{B})\ \left(\int_{\mathcal{Y}} p^g_{B|Y}(\overline{B}|\overline{Y})\ p_Y(\overline{Y})\ d\overline{Y}\right) d\overline{B}$$

[22] Note that this result does not follow from the law of iterated expectations. The law of iterated expectations can only be invoked in the fixed point. Outside the fixed point, $\mathcal{F}_{p_Y}(g)$ is not the marginal density of B consistent with p_Y and $p^g_{B|Y}$.

$$= \int_{\mathcal{Y}} \left(\int_{\mathcal{B}} h(\overline{B})\ p_g(\overline{B}|\cdot)\ d\overline{B} \right)\ p_Y = E_{p_Y}\ (\ E_{p^g(\cdot|Y)}(h(B))\ ) \qquad (17)$$

where the first equality follows by definition of $\mathcal{F}_{p_Y}(g)$, the second by Fubini and the third by definition of E_{p_Y}. This proves (14).

Clearly, (15) follows when we consider $h(B) = B$.

Also, (16) follows from (14) and

$$\mathrm{Var}_{\mathcal{F}_{p_Y}(g)}(B) = E_{\mathcal{F}_{p_Y}(g)}(B^2) - \left[E_{\mathcal{F}_{p_Y}(g)}(B)\right]^2 \qquad \square$$

This result immediately suggests the following Monte-Carlo approximation to compute $E_{\mathcal{F}_{p_Y}(g)}$ and $Var_{\mathcal{F}_{p_Y}(g)}$: draw M realizations of Y from p_Y; for each draw $\overline{Y}$ compute $E_{p^g(\cdot|\overline{Y})}(B)$ and $\mathrm{Var}_{p^g(\cdot|\overline{Y})}(B)$ using the closed-form expressions (12) and (13), finally approximate E_{p_Y} by averaging these closed-formed expressions to evaluate the right side of (15) and (16) over the M draws. The normal density with the resulting mean $E_{\mathcal{F}_{p_Y}(g)}(B)$ and variance $\mathrm{Var}_{\mathcal{F}_{p_Y}(g)}(B)$ is our proposed approximation to $\mathcal{F}_{p_Y}(g)$. This normal distribution can be interpreted as a second order approximation to $\log \mathcal{F}_{p_Y}(g)$.

In the empirical applications below we find approximate fixed points of $\mathcal{F}_{p_Y}$ by successive iterations. We start with a relatively flat normal distribution as an initial guess. We find successive means and variances of $\mathcal{F}_{p_Y}(g)$ using the approximate iteration described. We iterate until the scheme delivers satisfactory approximation to the desired prior marginal distribution of observables p_Y.[23] Obviously if such iteration failed to converge there are a number of search algorithms that could be used to find a fixed point of the mean and variance if the dimensionality of the problem is sufficiently low for gradient algorithms to work. The simplicity of successive approximations is highly desirable.

As with any algorithm it is important to check for accuracy. We should check that the approximate fixed point p_B satisfies (7) closely enough. This is even more important since, as of this writing, we do not have a sufficiency result for Proposition 1 stating that any fixed point of $\mathcal{F}_{p_Y}$ is indeed a prior consistent with p_Y. Checking for accuracy is straightforward: we draw parameter values B from the candidate fixed point, and then we simulate data for T_0 periods given this parameter value, drawing gaussian errors and starting from the initial observation in the data $\mathbf{x}_0 = (\mathbf{y}'_0, \mathbf{y}'_{-1}, ...\mathbf{y}'_{-P+1}, 1)$. This gives the distribution of Y in the left side of (7). We compare this

[23] At this writing we have no theorem that this algorithm will always work. However, in all practical applications we have tried, it delivered priors implying marginal data densities quite close to the desired one. In all cases it worked similarly as the analytically tractable special case in Appendix D: after the first few iterations the means of parameters stabilized, and subsequent iterations were only shrinking the prior variances.

distribution with the prior marginal distribution of the data p_Y to see if (7) approximately holds.

The approximate fixed point approach can be adapted to a wide range of problems. Linearity and normality are not essential for the algorithm to be feasible. What is key is a family of priors g and a likelihood for which $p^g_{B|Y}$ is known analytically. Any distribution for p_Y can be used, as long as we can generate random draws from it.

3.4 The case $T_0 = 1$, gaussian p_Y

In the case when $T_0 = 1$ the prior is only specified for the growth rate at the first date of the sample. When the prior about this growth rate is gaussian we can find an analytic expression for the delta prior:

Result 3. *In the VAR(P) model, given an initial condition* $x_0 \equiv (y_0', y_{-1}', ...y_{-P+1}', 1)'$ *assume the prior growth rate is*

$$\Delta y_1 | x_0 \sim N(\mu_\Delta, \Sigma_\Delta) \tag{18}$$

This is compatible with any prior p_B *satisfying*

$$B' x_0 \sim N(y_0 + \mu_\Delta, \Sigma_\Delta - \Sigma_u) \tag{19}$$

Note that, although our approach is to always set x_0 equal to the value $\mathbf{x_0}$ actually observed in the data, the above result holds for any value of the vector x_0. This generality will serve to make the connection with dummy observation priors in section 3.5.

Proof

We first show that (19) implies the mean and variance in (18). The VAR model implies

$$\Delta y_1 = B' x_0 + u_1 - y_0 \tag{20}$$

Taking expectations and using (19) we have $E(\Delta y_1) = \mu_\Delta$. Taking variances we have

$$var(\Delta y_1) = var(B' x_0) + \Sigma_u + cov(B' x_0, u_1') + cov(u_1, x_0' B)$$

Since the uncertainty about $B' x_0$ comes only from the uncertainty about parameters, these covariances are zero. Therefore using (19) we have

$$var(\Delta y_1) = \Sigma_\Delta - \Sigma_u + \Sigma_u = \Sigma_\Delta$$

Finally, normality of $B' x_0$ and u and (20) imply that Δy_1 is normally distributed.□

Comment 1: Prior (18) restricts N linear combinations of parameters. Therefore in general it needs to be completed with a prior on the remaining $K - N$ parameters where K is the total number of parameters.

Comment 2: The delta prior in section 2.4 follows from the above result using the actual observation in the data $\mathbf{x}_0$, and completing this prior with a flat prior in ρ.

Therefore, the advantage of prior (19) is that there is no need to engage in the numerical procedure that we discuss in section 3.3, since we have an analytic solution. The drawback is that taking $T_0 = 1$ means using very little prior information. In the empirical section 4.2 we find that using $T_0 > 1$ can make a big difference in terms of empirical results. For example, the "exact likelihood" approach amounts to introducing information about PN dimensions, which suggests setting $T_0 = P$ to introduce a similar amount of information.

Unfortunately, analytic solutions such as the one given in Result 3 are unlikely to be found when $T_0 > 1$, because then the prior about growth rates implies a complicated non-linear restriction on the distribution of shocks and parameters. To see this, consider the univariate AR(1) example of section 2. In this case, in addition to the prior about the first growth rate (3) one might specify a prior

$$\Delta y_2 | y_0 \sim N(\mu_\Delta, \Sigma_\Delta).$$

The above prior about Δy_2 implies

$$\alpha + (\rho - 1)\alpha + (\rho - 1)\rho y_0 + \rho u_1 + u_2 \sim N(\mu_\Delta, \sigma_\Delta)$$

in addition to (4). It should be clear that now the analytic solution is impossible. A change-of-variable formula can not be used to find the distribution of p_B because we cannot express the parameters α, ρ as a function of variables with a known distribution. This is due to the fact that the *joint* unconditional distribution of u's and y's consistent with the prior growth rates is non-trivial and unknown. More details are offered in Appendix C.

3.5 Relationship with dummy observation priors

We relate our approach with other Bayesian approaches in the context of specific applications in the next section.

The closest prior to ours is Sims' "one-unit-root" dummy observation prior, so this merits a more detailed discussion.[24] It is implemented by augmenting the sample with a "dummy observation" for an artificial date d. In

[24] See Sims and Zha (1998, eq.22 and Table 3) or Sims (2006, section 2.1).

this fictitious observation current and past values of the process are equal to $\lambda\bar{y}$, where λ is a constant specified a priori determining the weight given to the prior, and $\bar{y}$ is the mean of the initial observations $\bar{y} \equiv \frac{1}{P}\sum_{i=0}^{P-1} \mathbf{y}_{-i}$. So, the dummy observation is:

$$y_d = B'x_d + u_d \tag{21}$$

for $u_d \sim N(0, \Sigma_u)$ independent of U, $x_d = \lambda(\bar{y}', ..., \bar{y}', 1)'$ and $y_d = \lambda\bar{y}$. The literature has also used other dummy observation priors with different x_d. In absence of other prior knowledge, the posterior is computed by adding the dummy observation to the actual observations and using an otherwise flat prior.

It is easy to check that the posterior found in this manner is also for the posterior that would arise from stating the following prior about growth rates for $T_0 = 1$:

$$\Delta y_1 | (\bar{y}', \ldots \bar{y}', 1) \sim \mathrm{N}\left(0, (\lambda^{-2} + 1)\Sigma_u\right) \tag{22}$$

This follows from our Result 3 and inspection of the formula for the posterior.

Therefore, the one-unit-root prior is a special case of the prior about growth rates with four restrictions: First, the prior is restricted to the growth rate in one period only, it takes $T_0 = 1$. Second, the prior about growth rates is conditional on a particular fictitious state of the process given by $(\bar{y}', \ldots \bar{y}', 1)$, i.e. after P periods of no growth, while we use the actual initial condition $\mathbf{x}_0$ observed. Third, the mean growth rate is usually assumed to be zero. Fourth, the prior variance in (22) is equal to the variance of the error terms Σ_u increased by a factor $(\lambda^{-2} + 1)$.

As we have argued in section 3.4, the first restriction is binding: the dummy observation approach cannot be generalized to $T_0 > 1$, although extending T_0 introduces important information and it often matters in practice, as our applications below demonstrate. Setting $\lambda > 1$ does not correspond to a prior about growth rates for many periods, as for this case the numerical solution is needed.

The fourth restriction is binding too: when the variance of the growth rates prior Σ_Δ is not a scalar multiple of the error variance Σ_u, the prior cannot be implemented by adding a dummy observation, and instead our Result 3 should be used. In general there is no reason why the researcher's prior uncertainty about the growth rate of all variables is proportional to the variance of the innovations. Furthermore, the usual approach is to take $\lambda = 1$ and the standard interpretation is that this gives the prior the weight of one observation. But (22) shows that $\lambda = 1$ corresponds to the variance of the growth rate which is double that of the error term, which may or may not correspond with an actual prior about variables. This matters also the

empirical results: in the application in section 4.1 of this paper $\lambda = 1$ turns out to produce a very weak prior, leading to results that differ significantly from the results of our preferred prior.

The second and third restrictions can be generalized within the dummy observation approach and different values of λ can be set to change the impact of the prior on the posterior. However, it is not easy to elicit these elements of the prior in an intuitive way.

This brings us to the key difference between the priors about observables proposed in this paper and the dummy observations priors: Researchers interpret the dummy observations as "mental observations" on VAR parameters, and not on the observables themselves. But it is difficult to make convincing statements about VAR parameters, because they do not have an intuitive interpretation. In absence of such intuitive interpretation, any choice of e.g. the prior parameter λ seems arbitrary. Our approach allows the researcher to relate λ (or more generally Σ_Δ) to intuitive quantitative prior statements about observables.

The above discussion serves to rationalize the dummy observation approach and gives it a clear Bayesian interpretation. We think an advantage of our approach is that we are explicit about the interpretation of the prior about observables, which allows a meaningful elicitation of this prior. In addition, our approach allows the researcher to incorporate much more information by using $T_0 > 1$.

4 Empirical Applications

In this section we apply our approach in two empirical time series studies taken from the literature. The first is a univariate example. It is useful to demonstrate that available techniques can give a wide range of estimates with little guidance for choosing among them, while our approach provides a clear interpretation. The second example is a large-scale VAR. It shows that the algorithm we propose works in practice even in a case with many parameters and that it can make a difference for inference in practice.

The first step in the empirical analysis is to specify the prior about initial growth rates which should, in principle, be independent from the sample. Instead, we proceed in all cases by specifying the "prior" distribution based on the actual growth rates in the sample. Such data-based priors are common in the literature, they have well known shortcomings and advantages, so we do not comment on them any further. We do report sensitivity analysis to help the reader figure out the implications of her preferred Bayesian prior.

Our *baseline delta prior* is derived from a prior on growth rates that is

normally distributed with mean and standard deviation equal to the unconditional mean and standard deviation of growth rates in the sample. This prior conveys the assumption that the first few observations behave, in terms of growth rates, similarly as the rest of the sample.[25] We specify $T_0 = P$, the number of lags in the VAR, so that our prior carries as much information as the additional terms for the distribution of initial observations that would enter in the "exact likelihood". Finally, we assume that the variance matrix of growth rates is diagonal. This specifies the distribution for $N \times P$ observable variables. As the number of parameters is larger than $N \times P$, this prior is incomplete. We complete the prior with an approximately flat prior on the remaining dimensions of the parameter space.

We use the algorithm described in section 3.3. We start the iterations with a candidate prior which is normal with mean zero and variance equal to $10^4 I$ (where I is an identity matrix of appropriate dimension). In most cases after about 100-250 iterations densities of growth rates implied by the normal delta prior match very well the prior density on growth rates. By this time most variances have shrunk by many orders of magnitude, but some of them remain barely different from the starting point, consistently with the baseline delta prior being improper in some dimensions.

As in previous sections we assume for simplicity that the variance of shocks is known and we set it equal to the variance of OLS residuals from the analyzed autoregressive model.

4.1 Persistence of Stock Prices

In this subsection, we show the effect of the delta prior on the estimated persistence of Stock Prices measured by the log of the S&P500 index, observed annually from 1871 to 1988, taken from the Extended Nelson-Plosser (ENP) dataset (Nelson and Plosser, 1982; Schotman and Van Dijk, 1991). Many papers have tested for unit roots in this dataset. However, it has been argued that unit root tests usually have low power. Therefore, it is of interest to just characterize the uncertainty about long run properties of these series in terms of a posterior distribution, without reference to a particular point null hypothesis. The model used in these papers is AR(3) with intercept and trend:

$$y_t = \alpha + \gamma t + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \rho_3 y_{t-3} + u_t \tag{23}$$

[25] This assumption seems reasonable in our cases, but there are many situations where it would not be appropriate. For example, for a sample starting after the end of a war the researcher may want to specify a higher initial growth rate of GDP than if the sample started after a period of undisturbed growth. One could also use a training sample to inform this prior, but in our case earlier data is not available.

As Andrews and Chen (1994) we focus on the sum of the autoregressive parameters $\sum_{i=1}^{3} \rho_i$, which they argue is a relevant measure of persistence. It is straightforward to adapt our approach from section 3 to include the trend that appears in (23). Of course, we specify our prior on "total" growth rates, not on deviations from trend, since total growth rates are the quantity for which a prior distribution is natural to elicit.

The baseline prior about growth rates has mean 3.5% and standard deviation of 16%. Figure 5 illustrates the match between prior densities of growth rates and the densities of growth rates implied by the delta prior after 100 iterations. The solid line represents the left-hand side of equation (7) while the dashed line represents its right-hand side. It is clear from the picture that the match is very good, and that the normal approximation of section 3.3 works very well.

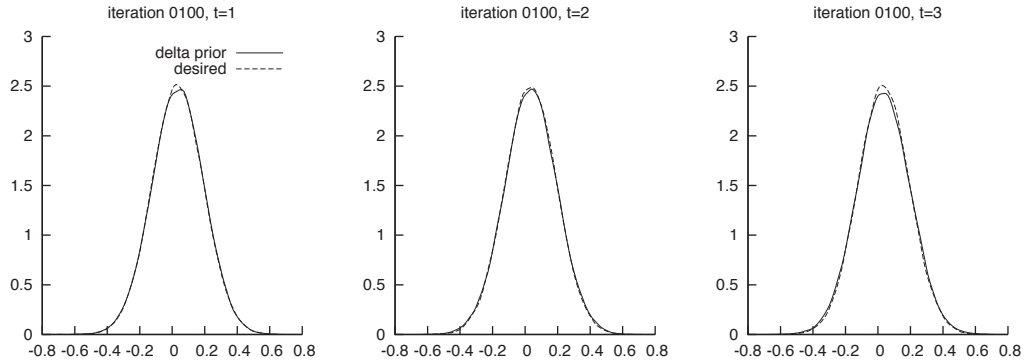


Figure 5 – Stock Prices, AR(3) with trend: density of growth rates in periods $t = 1, 2, 3$, obtained by Monte Carlo simulation; 'delta prior' - the marginal density of the data (growth rates) implied by the delta prior; 'desired' - the assumed growth rate which is to be matched by the delta prior.

Figure 6 compares the posterior found with the baseline delta prior with some other Bayesian and frequentist procedures that are currently available. We first discuss the other procedures available. We use the Minnesota prior with the standard hyperparameters recommended in Doan (2000). Perhaps surprisingly, this prior pushes persistence downward compared with the flat prior, even though it is centered at a unit root model! The reason is that the Minnesota prior shrinks the lagged parameters towards zero. This prior dampens the contributions of ρ_2 and ρ_3 to the persistence measure more than it pushes ρ_1 towards unity. One-unit-root dummy observation with $\lambda = 1$ has a very weak effect here and it delivers a similar persistence as the flat prior. As another comparison, we try the bootstrap-after-bootstrap

correction of the mean bias proposed for VARs by Kilian (1998).[26] This procedure produces a frequency distribution of the estimator, which is a different object than a posterior density, but we compare the two as is often done informally in applied work. Kilian's bias corrected estimation implies that the process is much more persistent than under all considered Bayesian procedures.

We also compare with the approximately median unbiased estimation of Andrews and Chen (1994). Andrews and Chen do not report the results as a density. Instead, they report the point estimate and a confidence interval. Their point estimate is 1, as a result of their assumed truncation of the parameter space to values $\left|\sum_{i=1}^{3} \rho_i\right| < 1$. Their initial estimate is larger than 1 and in this case they pull back the estimate to their upper bound of 1. The 90% confidence interval is $[0.91, 1]$ (see their Table 4 p.197), very similar to the 90% posterior probability interval obtained with the baseline delta prior.[27] However, the fact that the point estimate is 1 suggests that, overall, the Andrews and Chen approach yields stronger persistence than all other approaches except for the Killian's.

We conclude that the available Bayesian and frequentist techniques deliver a wide range of estimates of persistence. Point estimates/posterior means vary between 0.93 and 1, even though the sample spans over one century. These different estimates imply huge differences in the medium run behavior of the model. An applied economist would have a difficult time forming intuition about which procedure to choose.

Figure 6 also displays the posterior density of $\sum_i \rho_i$ using our baseline delta prior (drawn with the thick continuous line). As expected, given our discussion of section 2, the delta prior increases persistence and it pushes the estimate in the same direction as a classical bias correction. The increase in persistence relative to OLS (flat prior) is substantial ($E\sum_i \rho_i = 0.956$ with the delta prior, instead of 0.93 with OLS). This has strong effects on an impulse response function at medium lags. Our baseline delta prior suggests, in this case, that the posterior is between the flat prior (OLS) and the bootstrap bias-corrected estimates of persistence. To the extent that our prior is based on statements about observables which are easier to assess, we find it more convincing than the rules used to derive previously available procedures.

Figure 7 shows the sensitivity of the posterior to various choices of priors about growth rates. For the $T_0 = 1$ case the posterior is a bit more spread out, as is to be expected from a prior that uses less information, but the

[26] We do not restrict the polynomial in $\rho_1 ... \rho_3$ to be stationary. When we do shrink all nonstationary draws towards the unit root, as recommended by Kilian (1998), the density is simply truncated at 1 and has a spike there.

[27] The 5th and the 95th percentiles of the posterior are 0.90 and 1 respectively.

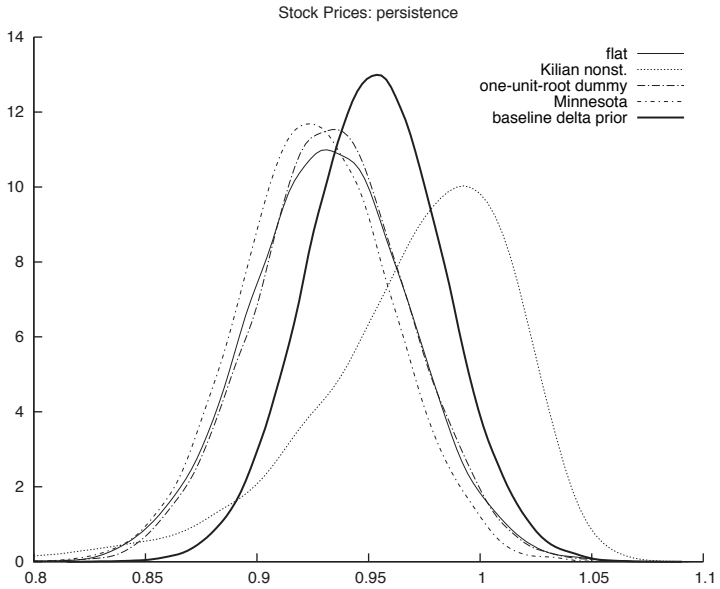


Figure 6 – Stock Prices, AR(3) with trend: density of the sum of autoregressive parameters $\sum_i \rho_i$. Various priors and estimation methods.

posterior mean is very close to that of the baseline prior. When we double the prior variance, so that the standard deviation of growth rate is 22.3%, the prior becomes very weak and the posterior (labeled "double variance") is very close to the flat-prior posterior. Arguably, this standard deviation is large, few people would argue this is a reasonable standard deviation of yearly growth rates of stock prices. This shows that it is important to use a "reasonable" variance of prior growth rates, one that does represent our opinion about the likely behavior of the series, in order for the prior to matter substantially for inference.

Next we model the prior in a richer way. Assuming no serial correlation in the prior growth rates as we have done in the baseline delta prior is questionable, because parameter uncertainty by itself implies a positive correlation in prior growth rates. Therefore we make an effort to use empirically-based serial correlation and to account for parameter uncertainty. For this purpose we use an auxiliary model. We fit an AR(1) model to growth rates $\Delta y_t = \alpha + \Gamma\, \Delta y_{t-1} + \varepsilon_t$ on the whole sample and we let our prior distribution of $(\Delta y_1, \Delta y_2, \Delta y_3)$ be given by this auxiliary model when the uncertainty on auxiliary parameters (α, Γ) is given by the flat-prior-posterior using the whole sample. To find this prior on observables we simulate the marginal density of growth rates in the first 3 periods, repeatedly drawing the parameters from the mentioned flat-prior-posterior. This density of growth rates

has a mean of 4.5%, standard deviation of 17.2%, correlation of consecutive growth rates of about 0.24 and the correlation between the first and the third growth rate of about 0.07. This prior has, therefore, both higher standard deviation and higher correlation than the baseline but the posterior is rather similar to the baseline case, as can be seen in the figure.

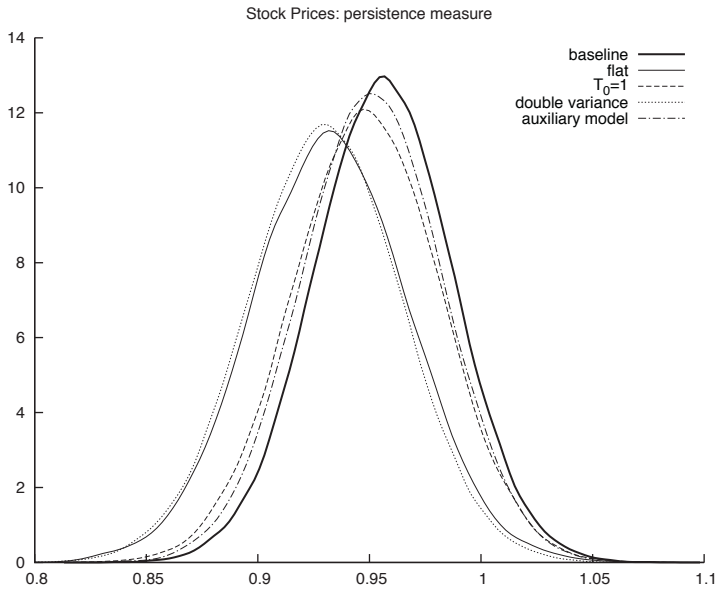


Figure 7 – Stock Prices, AR(3) with trend: Posterior density of the sum of autoregressive parameters $\sum_i \rho_i$. Various priors about initial growth rates.

Our conclusion is that reasonable informative priors give a similar picture: the estimated persistence of the S&P500 is corrected significantly upwards relative to OLS, but very little weight is placed on roots equal to or larger than 1.

4.2 Responses to Monetary Policy Shocks in a VAR

In this subsection we reconsider the estimation of the effects of a monetary shock in Christiano et al. (1999). They estimate a VAR with quarterly data on output, prices, commodity prices, federal funds rate, total reserves, non-borrowed reserves and money. (Details about data and sample are provided in Appendix E.) Residuals are orthogonalized with the Choleski decomposition of the variance of innovations given the above variable ordering. The monetary policy shock is the one corresponding to the federal funds rate.

Means and standard deviations of growth rates of the variables in the sample, which are used in the baseline prior, are reported in Table 1. Since

$T_0 = 4$ the dimension of the prior is only $4 \times 7 = 28$, compared with $4 \times 7^2 + 7 = 203$ parameters in the VAR, so the prior is quite weak.

After about 200 iterations the match between the 28 assumed densities of growth rates and their densities implied by the actually used gaussian prior is very good, we report these densities in Appendix E.

Figures 8 and 9 display the responses of output to a monetary policy shock, estimated with alternative approaches. For brevity, we focus on the response of output, first because it is a key policy issue, second because the output response is among the most affected by the frequentist small sample bias and by alternative prior assumptions. Responses of the remaining variables are reported in Appendix E.

Our benchmark is the posterior distribution of the impulse response obtained with the standard flat prior $p(B, \Sigma) \propto |\Sigma|^{-\frac{N+1}{2}}$. To facilitate comparisons we display this posterior as the shaded region in all plots. This is the region enclosed between quantiles 0.025 and 0.975 of the posterior distribution of the impulse response. The flat prior band is almost symmetric around the OLS point estimate.

Figure 8 compares the baseline posterior with other approaches used in the literature. The first plot reproduces the results in Christiano et al. (1999) who use a bootstrap procedure with the OLS point estimate of the parameters taken to be the data generating process. The continuous lines show the percentiles of the distribution of the impulse response estimated by OLS from the bootstrapped series. As is well known, the confidence bands from this procedure contain a second dose of OLS bias and, as a result, output response to the monetary policy shock dies out even sooner than under the flat prior.

Impulse response band obtained with our benchmark prior is displayed in the second plot of Figure 8. The effect on output is delayed considerably relative to the flat prior or the bootstrap procedure of Christiano et al. (1999). If the confidence bands from the delta prior show more persistence than the flat prior they contrast even more with the bootstrap bands which are less persistent than the flat prior. The effect on the economic interpretation of the results is fairly large. The cumulative effect of the shock after 4 years is -6.6% of the quarterly GDP (at the median) when estimated with the delta prior. This effect is only -4.6% when estimated with OLS and only -3.1% according to the bootstrap bands. *Therefore, the flat prior underestimates the cumulative effects of monetary shock by about one third, and bootstrap bands by more than a half, relative to the delta prior.*[28] Since, at least to

[28] Our conclusions are not affected by the fact that we condition the delta prior on the fixed value of σ_u^2 estimated by OLS. To verify this, we recomputed all impulse responses

our mind, the delta prior is a natural prior to impose, we conclude that the negative effects of monetary shocks on output have been previously underestimated. The difference is also large if we focus only on the short run (up to 6 quarters) since the delta prior implies that the effect on output has a larger delay than previously estimated.

The second row of plots in Figure 8 displays the effects of standard Bayesian VAR priors designed to push the persistence of the process upwards: the Minnesota prior and the Sims' one-unit-root prior. The Minnesota prior (with the default hyperparameters recommended by Doan (2000)) has little effect in this case and the impulse response band is close to the flat prior band. The one-unit-root prior (with $\lambda = 1$) increases the persistence of the response similarly to the delta prior, although it predicts a much larger negative effect in the short run.

The third row of Figure 8 displays, for comparison, the effect of applying frequentist bias correction procedures in the present context. The caveat mentioned in section 4.1 about comparing Bayesian posteriors and post-sample uncertainty applies here too. We show the bootstrap-after-bootstrap procedure to construct error bands for impulse responses proposed in Kilian (1998). We present results with two versions of this procedure: the one in which stationarity is imposed and the one in which we allow for nonstationarity. The results depend strongly on the handling of nonstationarity: in the first case the effect of monetary policy on output is much weaker than in the second case.

Again, it would be difficult for an applied economist to choose between the previously available alternatives and they do give different results. Our baseline results might be preferred since they are based on an intuitive and reasonable prior.

Figure 9 studies deviations from the various choices involved in the baseline delta prior. First, we consider the case $T_0 = 1$, when an analytical solution is available. The error bands almost overlap with the flat prior error band. Only when we increase to $T_0 = 4$ the effect of the prior kicks in, so the technical difficulties of translating the prior when $T_0 > 1$ are definitely worthwhile in this case. Second, we change the baseline prior assuming that $\mu_\Delta = 0$. Note that this prior is different from the one-unit-root prior in the previous Figure, because it has a different variance and assumes $T_0 = 4$. This prior makes the output responses more persistent (see the band labeled 'zero mean'). However, a zero mean for output growth rate is not a reasonable

in Figure 8 this time conditioning in all procedures on the fixed value of σ_u^2 estimated by OLS. The resulting bands were only slightly narrower and overall the results were very similar to those in Figure 8.

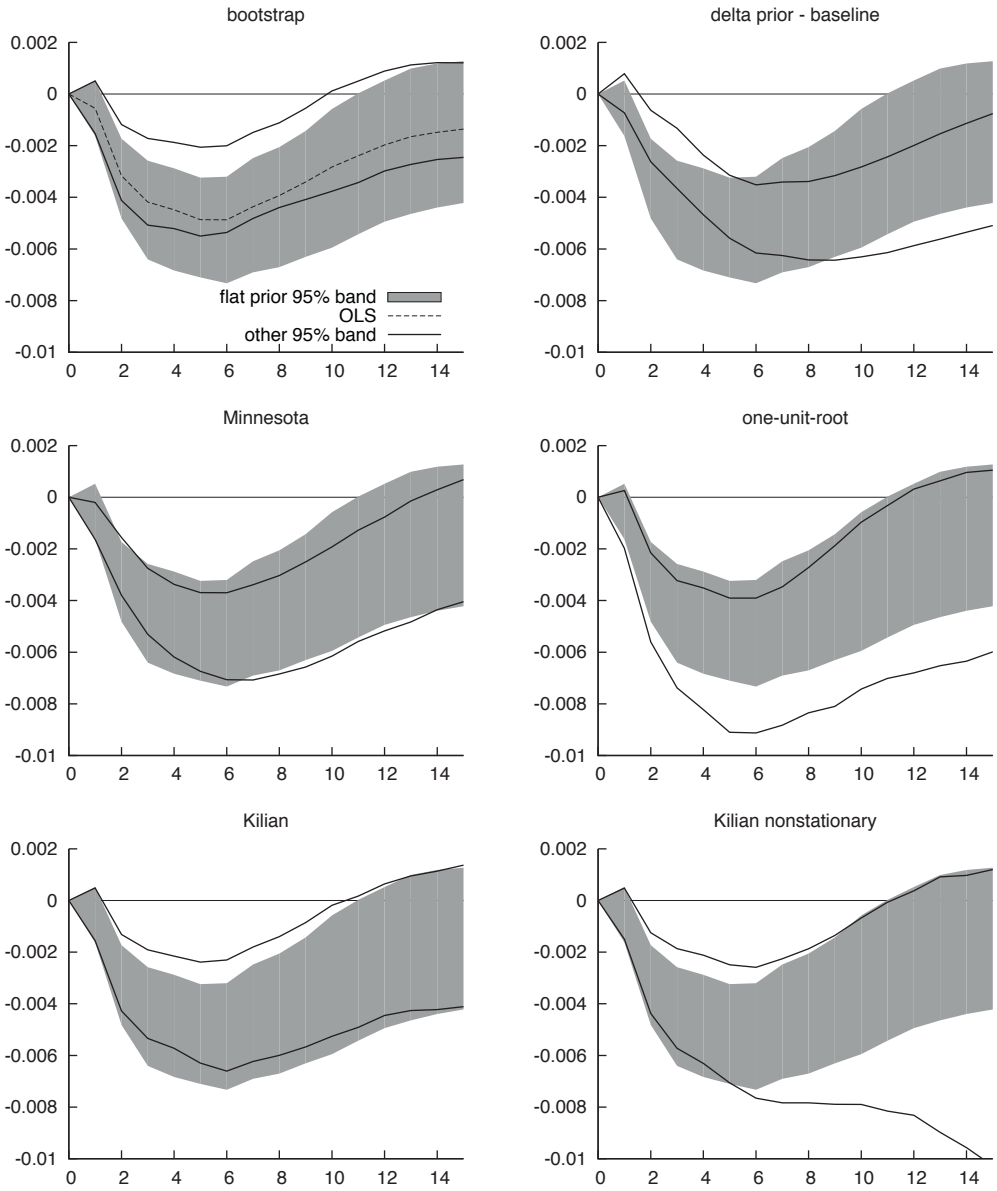


Figure 8 – Impulse responses of output to monetary shocks, 95% probability bands generated in alternative ways. In all plots the gray area shows the 95% probability band obtained with the flat prior.

prior. Finally, we use an auxiliary model as described in the study of Stock Prices to allow for a reasonable joint distribution of the growth rates. Output response band with this prior is slightly wider and less persistent than in the baseline case.

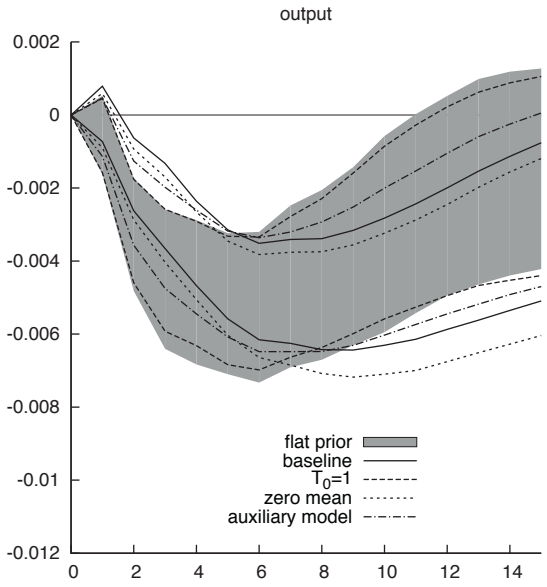


Figure 9 – Impulse responses of output to monetary policy shocks, 95% probability bands generated with various priors about initial growth rates.

We draw several conclusions from this example. First, it shows that our procedure for approximately solving equation (7) works efficiently in practice. With a standard personal computer it took about two hours to find a fixed point in the space of 20,909 parameters! (this includes the mean and the variance of the parameter matrix B, which has 203 entries). Second, imposing priors about initial growth rates in this case pushes the posterior estimates in the same direction as bias corrections. We find that the cumulative effects of monetary policy shocks on output have been strongly underestimated and they are more delayed than previously thought. There is little guidance as to what other available procedures should be chosen. For large models it may be important to specify $T_0 > 1$. Again, when specifying the delta prior in alternative but reasonable and informative ways the results are quite robust.

5 Frequentist Evaluation of a Delta Estimator in the AR(1) Model

In this section we study the adequacy of the delta prior in the AR(1) model from the classical point of view. We define a *delta estimator* to be the posterior mean obtained with the baseline delta prior. We find that, although this estimator is inspired by Bayesian principles it works very well under the usual Monte-Carlo evaluation procedures that classical econometricians use to justify the validity of small samples estimators. In fact, the delta estimator works better than classical procedures available.

Classical bias corrections have an element of arbitrariness in that a full correction of the mean bias is never achieved. In part for that reason, and in part because it is recognized that focusing on the bias is arbitrary, much of the bias correction literature ends up reporting the root mean squared error (RMSE) reduction for "relevant" parameter values as an important selling point of bias correcting estimators. We show that the delta estimator has a substantially lower RMSE than classical alternatives in a wide and empirically relevant range of parameter values.

We repeat the Monte Carlo study of MacKinnon and Smith (1998, section 5), adding the delta estimator to it. We simulate 100,000 realizations of the AR(1) process for each value of $\rho = 0.40, 0.42, 0.44, \ldots, 1.2$. This is a relevant range in many practical applications. In order to highlight the small sample problems we take a sample size $T = 25$. Initial observations are generated as in MacKinnon and Smith (1998).[29] For each realization of the process we estimate (α, ρ) with OLS, with the constant-bias-correcting (CBC) estimator of MacKinnon and Smith (1998) and with the delta estimator.

Figure 10 shows the biases of the three estimators for many values of ρ. The OLS estimator has the largest bias. The CBC estimator has a much smaller bias but, as is well known, the bias is not completely removed. We can see that for $\rho \in (0.68, 1.1)$ the bias of the delta estimator is in between that of OLS and CBC. Therefore, the Bayesian estimator also reduces the bias in this parameter range although the correction is less precise than CBC.

However, as is well known, bias reduction is not desirable per se, since it could lead to large RMSE. Figure 11 shows the RMSE of the three estimators. The CBC has a lower RMSE than OLS only when $\rho > 0.5$. But the delta estimator beats both OLS and CBC when $1.1 > \rho > 0.5$. When $\rho = 1$ the RMSE of the CBC is 21% larger than the RMSE of the delta estimator. Therefore, for roots close to unity the gain in efficiency of switching from CBC to the delta estimator is the same as if we suddenly found 50% more data

[29] We take $\alpha = 0$, so the initial condition amounts to $y_0 \sim N(0, \sigma_u^2)$ for all ρ.

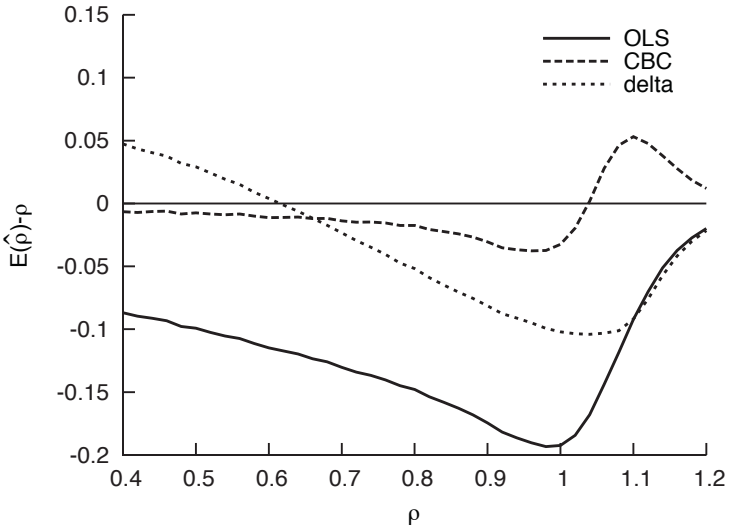


Figure 10 – Bias of the OLS, the CBC and the delta estimator, sample size T=25.

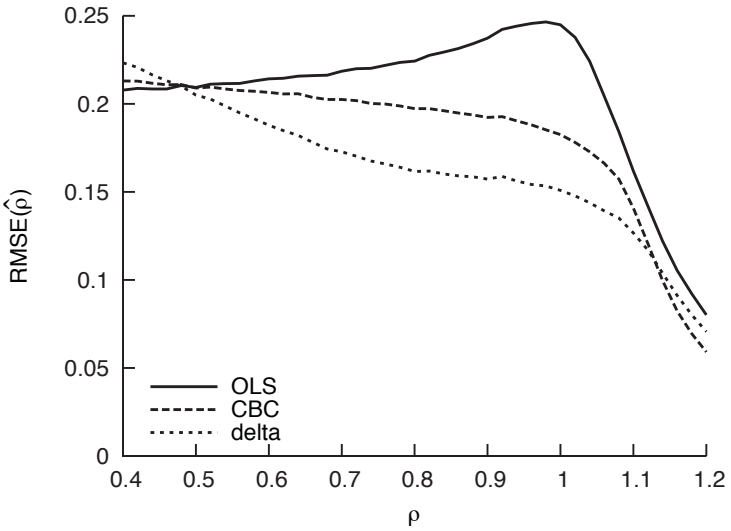


Figure 11 – RMSE of the OLS, the CBC and the two delta estimator, sample size T=25.

points and added them to the sample. We think this is a large improvement.

We have repeated the Monte Carlo study using other initial conditions from the literature and we obtained similar results. Notice that all the cards are stacked in favor of the CBC estimator, because the delta estimator uses only the current realization of the data while we always construct the CBC using the same initial condition as in the Monte Carlo study. Therefore, the CBC uses the information about the true initial condition used in the Monte-Carlo draws, in addition to each realization of the data. In the real world it is hard to know the "true" initial condition, so we would expect CBC to be at an even larger disadvantage in practice.

Our conclusion is that, as long as one is willing to believe that the true ρ is between .5 and 1.1, the delta estimator is an attractive alternative to bias corrections even from the frequentist point of view.

6 Conclusions

The estimation of autoregressions using small samples is a long-standing problem. Considerable effort has been devoted to this issue, often involving difficult analytical problems. A myriad of alternatives can be found in the literature designed to address this issue both from the Bayesian and classical point of view. But small sample issues are rarely addressed in practice. An applied economist has a hard time choosing among these alternatives because they require decisions that are not intuitive and, in practice, are often made ad hoc. Disappointingly, the most widely used alternative to estimate VARs is still OLS, which amounts to ignoring a problem that was pointed out sixty years ago.

We start by reexamining the classical versus flat-prior-Bayesian controversy about the validity of OLS. We find that for a similar treatment of initial conditions both Bayesian and classical econometricians agree that OLS should be adjusted towards a unit root. Therefore, what is important is to relate parameter values and observed initial conditions.

We propose to do this by specifying a prior about growth rates for the first few observations or, more generally, a prior about the behavior of the observables. Contrary to the contenders, this prior has a clear interpretation, it embodies information that economists do have, it is easy to elicit, and perhaps it is even possible seek a near-consensus about it.

Translating this prior about observables into a prior about coefficients we have to address a series of technical problems. Our approach to solving the relevant Fredholm equation may be useful in many other contexts.

To illustrate the effect of the delta prior we use it in two empirical ap-

plications from the literature. In a VAR for the US economy the delta prior delivers much more persistent response of output to monetary policy shocks than found in the literature. First, this shows that the delta prior matters in practice. Second, this illustrates that the tools developed in this paper allow to handle even large scale models like VARs. This opens a possibility of many more interesting empirical applications.

Even from the classical perspective, our Bayesian posterior estimates are attractive. Our estimator has an edge in terms of mean squared error relative to other classical bias correction procedures.

We conclude the delta prior is a way to approach the long standing problem of estimating autoregressions in small samples. Even though we advocate a Bayesian solution, our work also points to some dangers from the Bayesian approach, namely, that one can easily formulate priors on parameters that have totally unreasonable implications for series. From this point of view the flat prior is completely unreasonable, and using OLS in autoregressions is unwarranted. What is needed is a careful specification of informative priors, and specifying priors on observables is the more natural alternative.

Future research will no doubt improve the prior combining it further with other usable information from expert knowledge and available theoretical models, and it will sharpen the analytic results that we used to translate priors about observables into priors about parameters.

Appendices

Appendix A Construction of Figures 1, 2 and 4

Figure 1: Each density in Figure 1 is generated with the following Monte Carlo experiment. We simulate 20,000 realizations of the AR(1) process (1). For each realization we compute ρ^{OLS}. Then we approximate the frequentist density of ρ^{OLS} with a histogram of these 20,000 estimates.

In each simulation we take $\rho = 0.95$, sample length $T = 100$ and error variance $\sigma_u = 1$. We set $y_0 = 0$ and for each realization we draw α from the distribution consistent with (2) i.e. from $\mathrm{N}(0, \sigma_0^2/(\sum_{i=0}^{S-1} \rho^i))$, where $S = 100$. As discussed in the text, the only difference between the two densities is in the choice of σ_0^2.

We use a fixed $y_0 = 0$ and random α for consistency with the subsequent Bayesian experiments. At first glance this might seem inconsistent

with the frequentist discussion of Figure 1 which treats α as fixed and y_0 as random. However, the results in Figure 1 would have been exactly the same if we performed the frequentist simulation instead. To see this, suppose that instead we keep α fixed at an arbitrary assumed value and draw y_0 from $N\left(\alpha(\sum_{i=0}^{K}\rho^i), \sigma_0^2\right)$. Suppose that in this alternative simulation we use the same seed of the random number generator. Then it is easy to check that although in each draw we would obtain a different α^{OLS}, we would obtain the same ρ^{OLS} as in the corresponding draw of our baseline simulation.

Figure 2: We generated densities in Figure 2 following Sims and Uhlig (1991). That is, we perform a Monte Carlo simulation analogous to that underlying Figure 1 for each value of ρ on the grid 0.70, 0.71, ... 1.20. Then we line up the obtained histograms to obtain the bivariate density of ρ and ρ^{OLS}. Each density in Figure 2 is a cross-section of such bivariate density at $\rho^{OLS} = 0.95$. Therefore, it is the Bayesian posterior density of ρ conditional on a value of ρ^{OLS}. The prior underlying this posterior is

$$p(\alpha, \rho|y_0, y_{-S}, \sigma_u^2, \sigma_0^2) = p(\rho|y_0, y_{-S}, \sigma_u^2, \sigma_0^2)p(\alpha|\rho, y_0, y_{-S}, \sigma_u^2, \sigma_0^2) \tag{A.1}$$

where

$$p(\rho|y_0, y_{-S}, \sigma_u^2, \sigma_0^2) \propto 1d\rho \tag{A.2}$$

$$p(\alpha|\rho, y_0 = 0, y_{-S} = 0, \sigma_u^2, \sigma_0^2) = N(0, \sigma_0^2/(\sum_{i=0}^{S-1}\rho^i)) \tag{A.3}$$

The fact that the marginal prior for ρ is flat is reflected in the uniformly spaced grid of ρs in the Monte Carlo simulations. We verified that the truncation of the grid at 0.7 and 1.2 introduces only a negligible error, since, with the sample size $T = 100$, values of ρ beyond these bounds are quite unlikely to yield realizations that produce $\rho^{OLS} = 0.95$. The prior for α is implied by condition (2). σ_u^2 is a known constant equal to 1.

A question arises how sensitive Figures 1 and 2 are to various choices of parameter values. Let μ denote the deterministic component of y_0, i.e.

$$\mu = \alpha\left(\sum_{i=0}^{S-1}\rho^i\right) + \rho^S y_{-S} \tag{A.4}$$

The following results proves that the shape of the density of $\rho^{OLS}|\rho$ is invariant to the choice of μ and σ_u^2. As a consequence, the shape of the density of $\rho|\rho^{OLS}$ is also invariant to these choices.

Result 4. *Assume the model parameterized as*

$$y_t - \mu = \rho\,(y_{t-1} - \mu) + u_t \qquad \text{for } t = 1 \ldots T \tag{A.5}$$

and assume that the initial condition is given by:

$$y_0 = \mu + \sigma_u \psi \tag{A.6}$$

where ψ is a random variable. Then, if ψ independent of the shocks u and its distribution is independent of μ and σ_u the distribution of the OLS estimator of ρ in (1) is independent of μ and σ_u.

Proof. Define normalized errors: $v \equiv u/\sigma_u$. (A.6) allows to write:

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

where $\tilde{y}$ is the process with $\mu = 0$, 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:

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

Similar results about invariance of ρ^{OLS} have been used in the literature. Andrews (1993, Appendix A), contains a verbal proof for $|\rho| \leq 1$ and for a particular distribution for ψ. DeJong et al. (1992) contains a similar proof for a fixed initial displacement $y_0 - \mu$. As can be seen, the proof is very simple, but we could not find a formal result focused on giving a general form of the initial condition which guarantees independence of the distribution of ρ^{OLS} from nuisance parameters, so we offer it here for completeness.

Figure 4: The construction of the densities in Figure 4 is similar to the construction of densities in Figure 2, except that now to generate each realization of the process we draw α from (5). We take $\sigma_u = 0.057$, which is the standard error of an AR(1) model fitted by OLS to the 'Real GNP' series for the years 1909-1988, taken from the Extended Nelson-Plosser dataset of Schotman and Van Dijk (1991). σ_u matters here for the variance of the delta prior in (5) given the prior about growth rate.

Appendix B Proof of Result 1

We will use the notation $\sigma_\alpha^2 \equiv \sigma_\Delta^2 - \sigma_u^2$. Using a standard formula we know that the posterior mean conditional on σ_u^2 is

$$E\left(\begin{array}{c}\alpha\\ \rho\end{array}\middle| Y^T, \sigma_u^2\right) = \left(X'X + \left(\begin{array}{cc}\sigma_u^2\sigma_\alpha^{-2} & 0\\ 0 & 0\end{array}\right)\right)^{-1}\left(X'X\left(\begin{array}{c}\alpha^{OLS}\\ \rho^{OLS}\end{array}\right) + \left(\begin{array}{c}\sigma_u^2\sigma_\alpha^{-2}\mu_\alpha\\ 0\end{array}\right)\right) \tag{B.1}$$

where, taking $m_{kl} = \frac{1}{T}\sum_{t=1}^{T} y_{t-1}^k y_t^l$ for integer powers k, l

$$X'X = \left(\begin{array}{cc}T & Tm_{10}\\ Tm_{10} & Tm_{20}\end{array}\right)$$

Letting $\xi = \det\left(\begin{array}{cc}T + \sigma_u^2\sigma_\alpha^{-2} & Tm_{10}\\ Tm_{10} & Tm_{20}\end{array}\right)^{-1}$ it follows from simple algebra that

$$\left(X'X + \left(\begin{array}{cc}\sigma_\alpha^{-2} & 0\\ 0 & 0\end{array}\right)\right)^{-1} X'X = \left(\begin{array}{cc}\xi T^2\left(m_{20} - m_{10}^2\right) & 0\\ \xi T m_{10}\sigma_u^2\sigma_\alpha^{-2} & 1\end{array}\right).$$

Plugging this in (B.1) and using $\alpha^{OLS} = m_{01} - \rho^{OLS} m_{10}$ we have

$$E\left(\rho|Y^T, \sigma_u^2\right) = \rho^{OLS} + (m_{01} - \rho^{OLS} m_{10} - \mu_\alpha)\xi T m_{10}\sigma_u^2\sigma_\alpha^{-2}.$$

Simplifying, using $m_{01} = m_{10} + (y_T - y_0)/T$ and taking expectation over the posterior density of σ_u^2 we obtain that

$$E\left(\rho|Y^T\right) = \rho^{OLS} + \left[1 + \frac{(y_T - y_0)/T - \mu_\alpha}{m_{10}} - \rho^{OLS}\right]\xi T\left(m_{10}\right)^2 E(\sigma_u^2\sigma_\alpha^{-2}|Y^T). \tag{B.2}$$

Since $\xi > 0$ the whole term multiplying the brackets is positive. If $|(y_T - y_0)/T - \mu_\alpha| < \left(1 - \rho^{OLS}\right)|m_{10}|$ the bracket is positive, and this implies the result. □

Appendix C Our prior can not be found by a change of variable

It is sometimes more convenient to specify a prior about a nonlinear function of the parameters, rather than to specify a prior directly about the parameters. This amounts simply to reparameterizing the initial model. Villani (2009) uses this approach in an application related to the present paper. Alternatively, one can derive the implied prior about the original parameters

from the prior about their nonlinear function using the change of variable technique. This section shows that a prior about growth rates in a VAR cannot be handled with a change of variable. The reason is that growth rates are not a deterministic function of parameters. There is no one-to-one mapping between observables and parameters. In fact one can express parameters as a function of observables and shocks, (y, u). Therefore to apply the change of variable formula we would need the joint density of (y, u). This joint density would need to be consistent with the prior about observables, with the assumed density of the shocks, and with the independence of parameters and shocks. Unfortunately, the joint density of (y, u) satisfying these constraints is non-trivial and generally unknown. So this procedure does not work. To be more explicit, consider the AR(1) model with the constant term and $T_0 = 2$. In this case, the mapping from (α, ρ, u) to (y, u) is as follows:

$$y_1 = \alpha + \rho y_0 + u_1 \tag{C.1}$$
$$y_2 = \alpha + \alpha\rho + \rho^2 y_0 + \rho u_1 + u_2 \tag{C.2}$$
$$u_1 = u_1 \tag{C.3}$$
$$u_2 = u_2 \tag{C.4}$$

It is easy to verify that the Jacobian matrix of this transformation is:

$$\begin{pmatrix} 1 & y_0 & 1 & 0 \\ 1+\rho & \alpha + 2\rho y_0 + u_1 & \rho & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The determinant of this matrix is $\alpha + (\rho - 1)y_0 + u_1$, and the absolute value of this term multiplies the distribution in the new parameter space (α, ρ, u_1, u_2). This term cannot be factorized into terms involving only us and terms involving only the parameters. Therefore, the obtained density will not, in general, be consistent with independence of the model parameters and errors.

Appendix D Analytical iteration on the mapping $\mathcal{F}$ for AR(1)

Consider the AR(1) model without the constant term, with $y_0 \neq 0$ given. The first observation must be nonzero, because if we are unlucky to start exactly at the mean, growth rate in the first period does not depend on the parameters and the prior for growth rate will not carry any information about ρ. But we only require (7) to hold in a probability one set of ys. For

simplicity, σ_u^2 is given. Everywhere we will implicitly condition on y_0 and σ_u^2. The model is

$$y_t = \rho y_{t-1} + u_t \qquad u_t \text{ i.i.d. } \mathrm{N}(0, \sigma_u^2) \tag{D.1}$$

and the density of an observation in period 1 is

$$p_{y_1|\rho}(\bar{y}_1; \bar{\rho}) = N(\bar{\rho}\bar{y}_0, \sigma_u^2) \tag{D.2}$$

Introduce the prior assumption about zero (without loss of generality) growth rate in the first period:

$$p_{\Delta y_1}(\Delta \bar{y}_1) = \mathrm{N}(0, \sigma_\Delta^2) \tag{D.3}$$

which implies:

$$p_{y_1}(\bar{y}_1) = \mathrm{N}(y_0, \sigma_\Delta^2) \tag{D.4}$$

Let's find the marginal prior $p_\rho(\bar{\rho})$ which will be consistent with the above $p_{y_1|\rho}$ and p_{y_1}, i.e. which will satisfy

$$\int p_{y_1|\rho}(\bar{y}_1; \overline{\rho}) p_\rho(\overline{\rho}) d\bar{\rho} = p_{y_1}(\bar{y}_1) \tag{D.5}$$

D.1 Guess of the solution

It is easy to guess that the solution is:

$$p_\rho^{guess}(\bar{\rho}) = \mathrm{N}\left(1, \frac{\sigma_\Delta^2 - \sigma_u^2}{y_0^2}\right) \tag{D.6}$$

Verifying (we skip algebraic details which are tedious, the integral can be performed by completing the square):

$$\int p_{y_1|\rho}(\bar{y}_1; \bar{\rho}) p_\rho^{guess}(\bar{\rho}) d\bar{\rho} = ... = (2\pi)^{-\frac{1}{2}} \sigma_\Delta^{-1} \exp\left(-\frac{1}{2}\frac{(\bar{y}_1 - y_0)^2}{\sigma_\Delta^2}\right) = p_Y(\bar{y}_1) \tag{D.7}$$

so the guess was right: $p_\rho^{guess}(\bar{\rho})$ satisfies condition (D.5).

D.2 Approaching the prior by fixed point iteration

Suppose we start with the flat prior $p(\rho) \propto 1$. One iteration with mapping $\mathcal{F}$ produces:

$$p_\rho^{\mathcal{F}(1)}(\bar{\rho}) = \int \frac{p(\bar{y}_1; \bar{\rho}) \times 1}{\int p(\bar{y}_1; \widetilde{\rho}) \times 1 d\widetilde{\rho}} p_Y(\bar{y}_1) d\bar{y}_1 = ... = \mathrm{N}\left(1, \frac{\sigma_\Delta^2 + \sigma_u^2}{y_0^2}\right) \tag{D.8}$$

As before, the integral is tedious but easy to compute by 'completing the square'. Verifying if $p_\rho^{\mathcal{F}(1)}$ satisfies D.5, i.e. if it is consistent with the desired marginal distribution of y_1 yields:

$$\int p_{y_1|\rho}(\bar{y}_1;\bar{\rho})p_\rho^{\mathcal{F}(1)}(\bar{\rho})d\bar{\rho} = ... = \mathrm{N}\left(y_0, \sigma_\Delta^2 + 2\sigma_u^2\right) \neq p_Y(\bar{y}_1) \tag{D.9}$$

The marginal distribution of y_1 implied by $p_\rho^{\mathcal{F}(1)}(\bar{\rho})$ is not what we wanted. It has the correct mean, but the variance is too high. In the second iteration, first we compute the prior $p_\rho^{\mathcal{F}(\mathcal{F}(1))}(\bar{\rho})$ by applying mapping $\mathcal{F}$ to the prior obtained in the first step

$$p_\rho^{\mathcal{F}(\mathcal{F}(1))}(\bar{\rho}) = \int \frac{p(\bar{y}_1;\bar{\rho}) \times p_\rho^{\mathcal{F}(1)}(\bar{\rho})}{\int p(\bar{y}_1;\widetilde{\rho}) \times p_\rho^{\mathcal{F}(1)}(\widetilde{\rho})d\widetilde{\rho}} p_Y(\bar{y}_1)d\bar{y}_1 = \dots$$

$$\cdots = \mathrm{N}\left(1, \frac{\sigma_\Delta^2 + \sigma_u^2}{y_0^2} \times \frac{\sigma_\Delta^4 + 2\sigma_\Delta^2\sigma_u^2 + 2\sigma_u^4}{(\sigma_\Delta^2 + 2\sigma_u^2)^2}\right) \tag{D.10}$$

Conveniently, we already computed the integral in the denominator while verifying $\mathcal{F}(1)$ (equation D.9 above). This prior has a smaller variance than the prior from the first step. To see this, note that the second quotient in the variance is less then 1, which can be seen after expanding the denominator. So the prior $\mathcal{F}(\mathcal{F}(1))$ has a smaller variance then the prior $\mathcal{F}(1)$. However, it still does not satisfy (D.5):

$$\int p_{y_1|\rho}(\bar{y}_1;\bar{\rho})p_\rho^{\mathcal{F}(\mathcal{F}(1))}(\bar{\rho})d\bar{\rho} = ... = \mathrm{N}\left(y_0, \frac{\sigma_\Delta^6 + 4\sigma_\Delta^4\sigma_u^2 + 8\sigma_\Delta^2\sigma_u^4 + 6\sigma_u^6}{(\sigma_\Delta^2 + 2\sigma_u^2)^2}\right) \neq p_Y(\bar{y}_1) \tag{D.11}$$

The marginal distribution of y_1 implied by $p_\rho^{\mathcal{F}(\mathcal{F}(1))}(\bar{\rho})$ is still not right. The mean remains correct. The variance is smaller than in the first step, but larger than the correct variance.

$$\sigma_\Delta^2 < \frac{(\sigma_\Delta^2 + 2\sigma_u^2)^3 - (2\sigma_\Delta^4\sigma_u^2 + 4\sigma_\Delta^2\sigma_u^4 + 2\sigma_u^6)}{(\sigma_\Delta^2 + 2\sigma_u^2)^2} < \sigma_\Delta^2 + 2\sigma_u^2 \tag{D.12}$$

The transformation is intended to facilitate seeing the second inequality. The first inequality is easy to prove too. Concluding, in the second iteration we got closer to the right prior.

Appendix E Data and additional results for the monetary VAR

The data for the Christiano et al. (1999) VAR were downloaded from Christiano's webpage. All data are quarterly and the sample is from 1965Q3

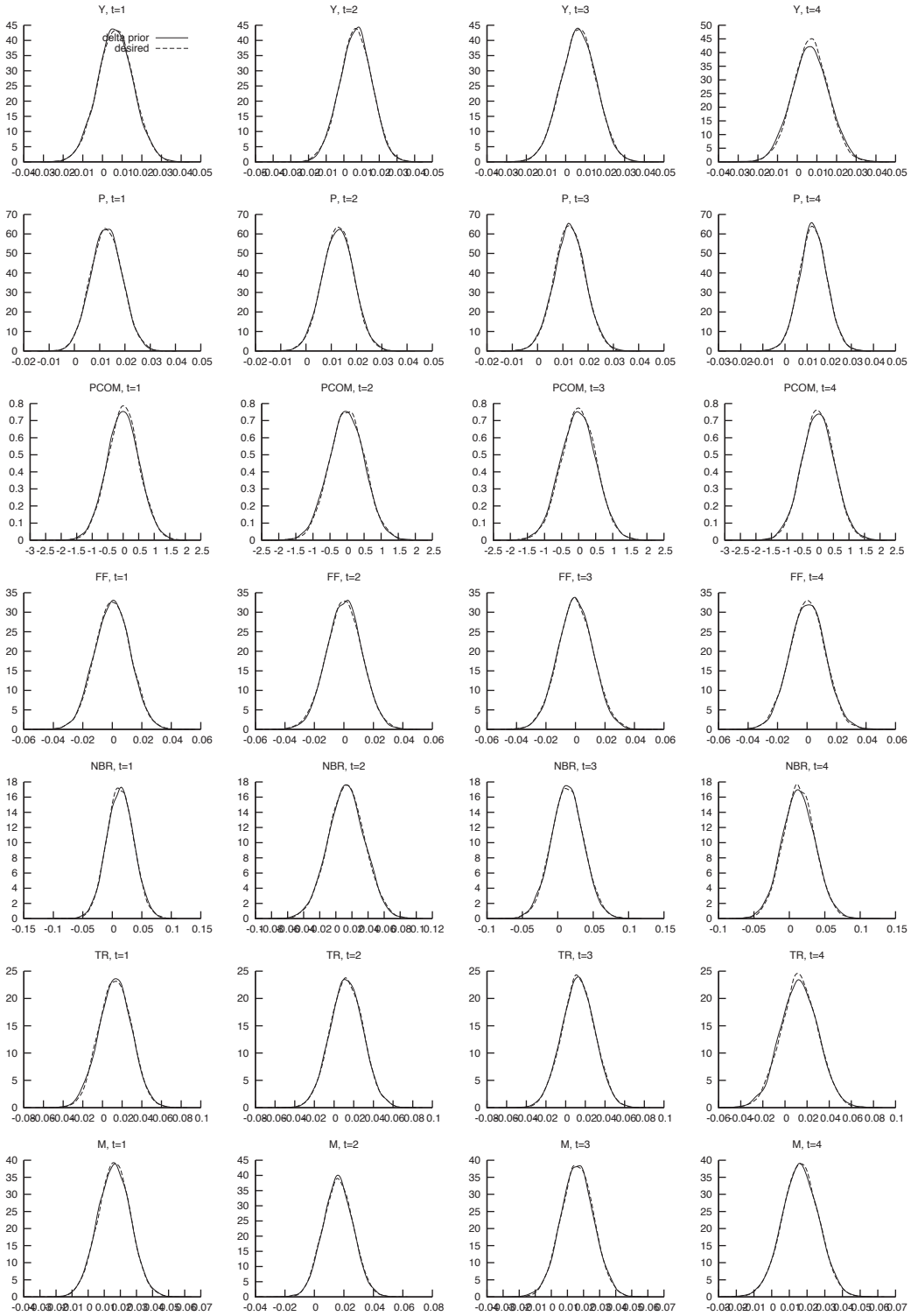


Figure 12 – Densities of growth rates of all variables in the periods t=1,2,3,4.

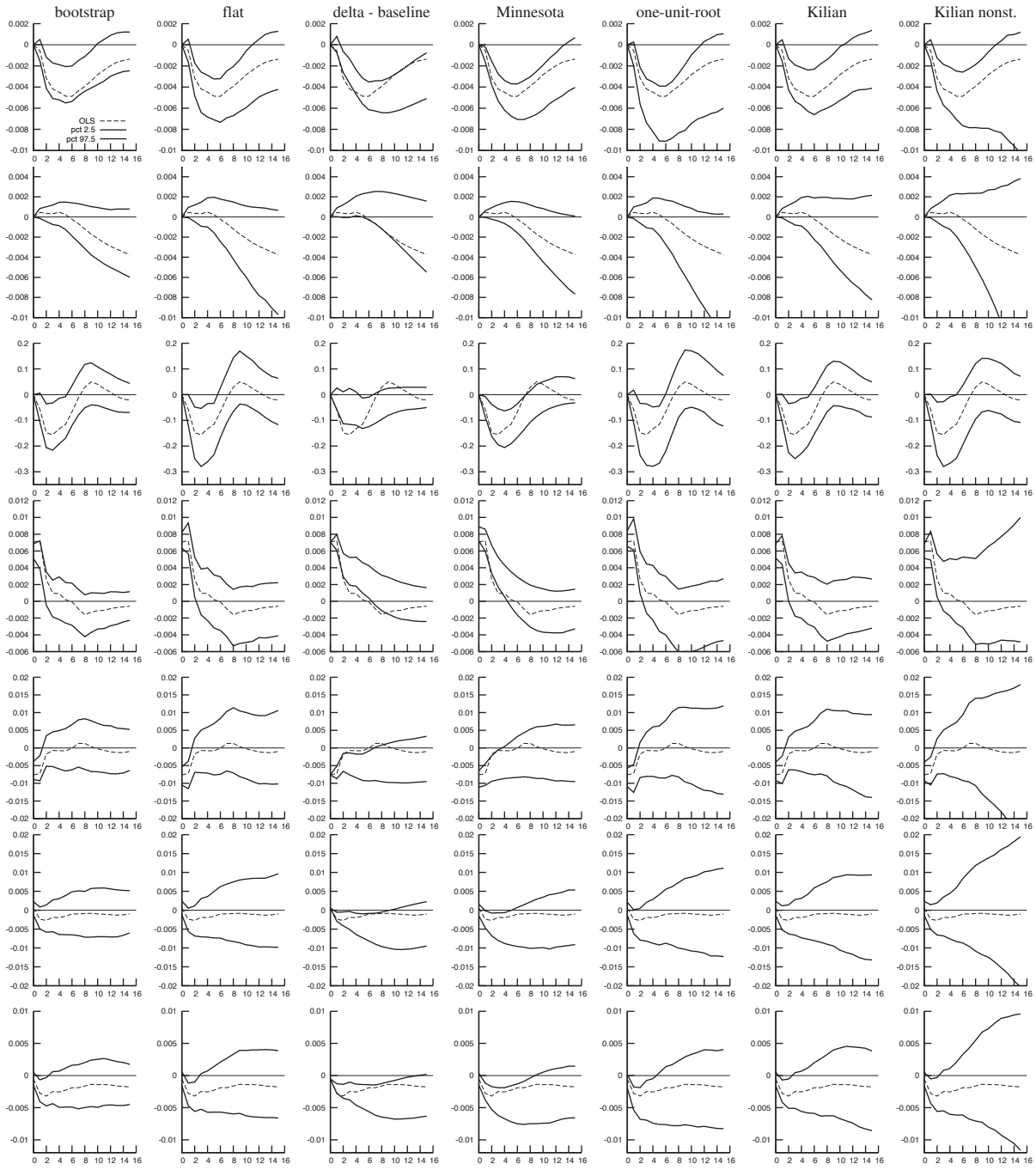


Figure 13 – Impulse responses to monetary shocks: OLS point estimate (dashed line) and the 95% uncertainty bands (continuous lines) generated by alternative methods

to 1995Q2. Table 1 reports means and variances of their first differences. Figure 12 reports the match between the prior densities of growth rates and

Table 1 – Average growth rates and standard deviations of the endogenous variables in the sample (1965Q3:1995Q2)

variable	definition	mean annualized growth rate	annualized standard deviation
Y	real GDP, logs	2.7	3.6
P	implicit GDP deflator, logs	5.0	2.5
PCOM	smoothed change in an index of sensitive commodity prices	3.2	206
FF	Federal Funds rate	0.1	4.8
NBR	nonborrowed reserves, logs	5.4	9.1
TR	total reserves, logs	5.2	6.6
M1	M1, logs	6.5	4.0

Note: The quarterly growth rates and their standard deviations are multiplied by 4. The original quarterly values were used in the prior.

the densities of growth rates implied by the delta prior. Figure 13 reports impulse responses of all variables to the monetary policy shock. In each plot, continuous lines delimit the 95% probability band. The OLS point estimate is also plotted on each plot for comparison, with the dashed line.

References

Abadir, K. M., Hadri, K., and Tzavalis, E. (1999). The influence of VAR dimensions on estimator biases. *Econometrica*, 67(1):163–181.

Andrews, D. W. K. (1993). Exactly median-unbiased estimation of first order autoregressive / unit root models. *Econometrica*, 61(1):139–165.

Andrews, D. W. K. and Chen, H.-Y. (1994). Approximately median-unbiased estimation of autoregressive models. *Journal of Business and Economic Statistics*, 12(2):187–204.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press, first edition.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, New York, second edition.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*, volume 6 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, Hayward, California, second edition.

Bhargava, A. (1986). On the theory of testing for unit roots in observed time-series. *Review of Economic Studies*, 53(3):369–384.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115 – 143.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Chapter 77 linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. volume 6, Part 2 of *Handbook of Econometrics*, chapter 77, pages 5633 – 5751. Elsevier.

Chamberlain, G. (2000). Econometrics and decision theory. *Journal of Econometrics*, 95(2):255 – 283.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? In Taylor, J. B. and Woodford, M., editors, *Handbook of Macroeconomics*, number 1A, chapter 2, pages 65–148. Amsterdam: North-Holland.

DeJong, D. N., Nankervis, J. C., Savin, N. E., and Whiteman, C. H. (1992). Integration versus trend stationary in time series. *Econometrica*, 60(2):423–433.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projections using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Doan, T. A. (2000). *RATS version 5 User's Guide*. Estima, Suite 301, 1800 Sherman Ave., Evanston, IL 60201.

Hurwicz, L. (1950). Least-squares bias in time series. In Koopmans, T. C., editor, *Statistical Inference in Dynamic Economic Models*. Wiley, New York.

Kadane, J. B., Chan, N. H., and Wolfson, L. J. (1996). Priors for unit root models. *Journal of Econometrics*, 75(1):99–111.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75(372):845–854.

Kendall, M. G. (1954). Note on the bias in the estimation of autocorrelation. *Biometrika*, XLI:403–404.

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, 80(2):218–230.

Lubrano, M. (1995). Testing for unit roots in a bayesian framework. *Journal of Econometrics*, 69:81–109.

MacKinnon, J. G. and Smith, A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, 85:205–230.

Marriott, F. H. C. and Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika*, XLI:393–403.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica*, 75(5):1411–1452.

Müller, U. K. and Elliott, G. (2003). Tests for unit roots and the initial condition. *Econometrica*, 71(4):1269–1286.

Nelson, C. R. and Plosser, C. R. (1982). Trends and random walks in macroeconmic time series : Some evidence and implications. *Journal of Monetary Economics*, 10(2):139 – 162.

Orcutt, G. H. and Winokur, H. S. (1969). First order autoregression: Inference, estimation, and prediction. *Econometrica*, 37(1):1–14.

Phillips, P. C. B. (1987). Time series regression with a unit root. *Econometrica*, 55(2):277–301.

Phillips, P. C. B. (1991). To criticize the critics: An objective bayesian analysis of stochastic trends. *Journal of Applied Econometrics*, 6(4):333–364.

Phillips, P. C. B. and Magdalinos, T. (2009). Unit root and cointegrating limit theory when initialization is in the infinite past. *Econometric Theory*, 25(06):1682–1715.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society Series B*, 11:68–84.

Roy, A. and Fuller, W. A. (2001). Estimation for autoregressive time series with a root near 1. *Journal of Business and Economic Statistics*, 19(4):482–493.

Schotman, P. C. and Van Dijk, H. K. (1991). On bayesian routes to unit roots. *Journal of Applied Econometrics*, 6(4):387–401.

Sims, C. A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of Econometrics*, 95:443–462.

Sims, C. A. (2006). Conjugate dummy observation priors for VAR's. Technical report, Princeton University.

Sims, C. A. and Uhlig, H. (1991). Understanding unit rooters: A helicopter tour. *Econometrica*, 59(6):1591–1599.

Sims, C. A. and Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, 39(4):949–68.

Stine, R. A. and Shaman, P. (1989). A fixed point characterization for bias of autoregressive estimators. *The Annals of Statistics*, 17(3):1275–1284.

Uhlig, H. (1994a). On Jeffreys prior when using the exact likelihood function. *Econometric Theory*, 10(3-4):633–644.

Uhlig, H. (1994b). What macroeconomists should know about unit roots - a bayesian perspective. *Econometric Theory*, 10(3-4):645–671.

Villani, M. (2009). Steady state priors for vector autoregressions. *Journal of Applied Econometrics*, 24(4):630–650.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.