Priors about Observables in Vector Autoregressions*

Marek Jarociński
European Central Bank

Albert Marcet
University College of London,
Barcelona GSE and CEPR

December 28, 2018

Abstract

Standard practice in Bayesian VARs is to formulate priors on the autoregressive parameters, but economists and policy makers actually have priors about the behavior of observable variables. We show how to translate the prior on observables into a prior on parameters using strict probability theory principles, a posterior can then be formed with standard procedures. We state the

*We thank Gianni Amisano, Manolo Arellano, Stéphane Bonhomme, Tony Braun, Matteo Ciccarelli, Jean-Pierre Florens, Bartosz Maćkowiak, Leonardo Melosi, Ricardo Reis, Juan Rubio-Ramirez, Thomas J. Sargent, Frank Schorfheide, Chris Sims, Jim Stock and Harald Uhlig for their comments and Atanas Pekanov for excellent research assistance. All errors are our own. Albert Marcet acknowledges support from Programa de Excelencia del Banco de España, Axa Research Fund, Plan Nacional (Ministerio de Educación y Ciencia), SGR (Generalitat de Catalunya), Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Program for Centers of Excellence in R&D (SEV-2011-0075) and European Community FP7-SSH grant MONFISPOL under grant agreement SSH-CT-2009-225149. The opinions expressed herein are those of the authors and do not necessarily represent those of the European Central Bank. Contacts: marcet.albert@gmail.com and marek.jarocinski@ecb.int.

inverse problem to be solved and we propose a numerical algorithm that works well in practical situations. We prove equivalence to a fixed point formulation and a convergence theorem for the algorithm. We use this framework in two well known applications in the VAR literature, we show how priors on observables can address some weaknesses of standard priors, serving as a cross check and an alternative formulation.

Keywords: Bayesian Estimation, Prior Elicitation, Inverse Problem, Structural Vector Autoregression

JEL codes: C11, C22, C32

1 Introduction

The application of Bayesian methods has been a key element in the development of vector autoregressions (VARs) and it has allowed for much progress in their application.[1] The literature offers a variety of priors on VAR parameters, from a practical point of view it is difficult to know which prior is the appropriate one in a given application and the choice of the prior often matters significantly for the results.

From a strictly Bayesian point of view the fact that different priors give rise to different posteriors is not necessarily a problem. If a prior on parameters really represents the beliefs of the analyst, the resulting posterior gives the correct answer for these prior beliefs. In this case different posteriors would appropriately reflect differences in prior beliefs. However, VAR parameters usually lack intuitive interpretation so it is difficult to claim that an analyst has genuine prior beliefs about VAR parameters.[2] This is an important stumbling block in the Bayesian analysis of VARs.

[1]VARs in macroeconomics follow from Sims (1980). See Rubio-Ramírez et al. (2010) on the identification of structural VARs, and Sims and Zha (1998) on Bayesian VARs.

[2]To be specific: an analyst estimating the mean of a population, or the elasticity of substitution

We propose to formulate the prior on the observables instead of the VAR parameters. To the extent that economists do have priors about the behavior of observable time series our proposal is to be 'truly Bayesian' and to incorporate this prior knowledge in the estimation of VARs. This can be done by, first, 'translating' the prior on observables to an equivalent prior on parameters and then, obtaining the posterior in the usual way. We do the 'translation' by solving an inverse problem, a Fredholm equation of the first kind. We propose an algorithm to solve this equation by reformulating this inverse problem as the fixed point of a certain mapping. We prove that under mild assumptions this fixed point condition is necessary and sufficient for the solution and that an algorithm based on successive approximations converges locally to the fixed point. Finally, we propose an approximate conjugate algorithm that speeds up computation.

To show that our approach works in practical applications we use it to reexamine two important VAR studies: the estimation of fiscal policy effects in Blanchard and Perotti (2002) and the estimation of monetary policy effects in Christiano et al. (1999). In each case we use a subjective prior about observables and compare with a few most popular variants of the standard priors for VARs due to Litterman (1979); Sims and Zha (1998). These examples serve several purposes. First, to show that different standard priors available in the literature can give very different results, an applied economist would have a hard time choosing among them. Second, that these standard priors on parameters imply very disparate behavior of observables and sometimes prior beliefs on observables that a reasonable analyst would not hold, hence in such cases the implied posterior is not well grounded on subjective Bayesian

between two goods, can have a subjective prior about the mean and the elasticity because these parameters have an intuitive interpretation. But it is difficult to give an interpretation, say, to the coefficient of the third lag of GDP in the VAR equation with the price level on the left hand side, so an analyst is unlikely to have a subjective prior about it.

principles. Third, that the algorithm we propose works in these examples and it gives an accurate solution to the inverse problem. Fourth, we demonstrate different methods to set up priors on observables.

It is of independent interest that priors on observables matter for the economic implications of these empirical studies. With our priors on observables we find a government spending multiplier that is about 50% larger than in Blanchard and Perotti (2002) and the real effects of monetary policy that are almost twice as persistent as in Christiano et al. (1999), resulting in a 30% larger cumulative effect on real GDP after 5 years. The priors on observables help clarify empirical results, as they eliminate some of the inconsistencies that priors on parameters generate. Also, they reduce the posterior variance relative to the noninformative prior by incorporating useful information into the inference, and since the posterior is derived from a proper subjective prior the results have a clear Bayesian interpretation.

Differently from our approach, most papers on priors for VARs provide a recipe for constructing the prior. Litterman (1986); Sims and Zha (1998) state rules of thumb for specifying a few prior hyperparameters that determine the prior distribution. Giannone et al. (2015) estimate these hyperparameters. Del Negro and Schorfheide (2004) generate a prior for a VAR linked to a DSGE model and estimate both the DSGE model and the VAR simultaneously. By contrast, we offer no general recipe. Our priors about observables are subjective and application specific, which, of course demands more from the user. In this paper we do not take a stand on what is the best way to specify a prior on observables, we merely point out that there are various methods to do so, and to make this point clear we use a different method for each example: we use knowledge about the economy stated by Blanchard and Perotti in the first example and an empirical Bayes prior in the second example. Future research should be directed at practical and intuitive ways of specifying priors about

observables.

Our work gives a different perspective on the interpretation of the existing priors. At least, researchers should examine if the VAR prior on parameters that they use implies a reasonable prior behavior for observables, as can be done with the accuracy check that we propose in section 3.3. In some cases the standard Sims and Zha priors can imply reasonable behavior of the observables, and hence could be used as simplified priors on observables. In other cases they include dynamics of the observables that an economist would rule out. In any case, the flat prior, used in some Bayesian VARs and implicit in the frequentist VARs, is even worse, it always implies crazy beliefs about the observables. For example, it would imply a prior statement that GDP is very likely to grow by more than, say, 100% in one quarter, a prior belief that no economist would hold. In macroeconomic applications samples are often short and priors matter. Consider an applied economist having to choose from a menu of standard priors in VARs. From our perspective this economist should at least check which prior has the most reasonable implications for observables. From this vantage point the flat prior is likely to have a rough ride. Even better if this researcher had enough time to specify his own prior on observables and apply our methodology for translating priors.

What constitutes a good prior in practice is a very complicated issue. Since Litterman (1979) it is customary to judge the practical virtues of VAR priors by examining their out-of-sample forecasting performance. The prior of Sims and Zha (1998) is advocated on those grounds. Further research is needed on how the 'reasonableness' of the priors on observables translates into improvements in forecasting performance. Another issue is the frequentist evaluation of priors on observables.[3] In this paper

[3]Jarociński and Marcet (2010) show an example of an empirical Bayes prior on observables that reduces the mean squared error of the estimator in the autoregressive model relative to the various classical small sample bias correction techniques considered.

we focus on finding a posterior taking for granted that a certain prior on observables represents the analyst prior knowledge, we leave the study of the above practical issues for future research.

Section 2 states the problem of mapping a prior on observables into prior on parameters, section 3 presents the fixed point formulation of this problem and convergence theorems, section 4 shows the empirical applications. The appendix contains the proofs and details of the empirical applications. An appendix available online provides additional implementation details, empirical and Monte Carlo results.

Related literature

Prior elicitation. Almost all applications in Bayesian econometrics are based on priors specified directly on parameters, and not on observables. Kadane et al. (1980) and Berger (1985, Ch.3.5) advocate specifying priors on observables, but they acknowledge the difficulty of solving the inverse problem in practice and their recommendation has had limited impact in econometrics. Kadane et al. (1996) is a small scale time series application.

Priors for VAR parameters used in the literature are loosely motivated by the implied behavior of the series. Such motivations stand behind the Litterman, Sims and Zha priors (Litterman, 1979, and others), steady-state priors (Villani, 2009), priors about the cointegrating relations in the data (Giannone et al., 2018), DSGE model-based priors (Ingram and Whiteman 1994, Del Negro and Schorfheide 2004, Del Negro et al. 2007, Christiano et al. 2011 and others) or 'system priors' of Andrle and Plasil (2016). However, in most of these approaches the prior information on observables is stated informally, and the connection between the prior on parameters and on observables is also informal. Our paper is the first to derive a VAR posterior from a prior on observables applying strict probability theory.[4]

[4]For example, the DSGE-model-based priors or 'system priors' in effect do not solve the inverse

Numerical methods. Inverse problems have attracted interest in microeconometrics recently, see Carrasco et al. (2007) for a survey. This literature focuses on issues of consistency and asymptotic distribution while we are interested in the computation of a prior on parameters. More importantly, the numerical methods used in this literature would be unfeasible for the high-dimensional problems that we face.[5]

One common theme in the literature just mentioned is whether or not a solution exists and the inverse problem is well-posed. We do not focus on these issues in this paper. The analyst can check ex-post if the solution to our fixed point problem implies a density of observables that captures approximately his prior, alleviating the problem of existence. We discuss these issues in detail in section 4 in the context of the three applications we consider. Furthermore, the approximate conjugate algorithm that we use appears to act as a 'regularization' of the kind that is often used in inverse problems to go around the numerical difficulties that are encountered in ill-

problem described in section 2. In light of our results, they can be justified as performing one iteration on the mapping on which we find one should iterate until finding the fixed point.

[5]To mention two recent papers in this literature. Bonhomme and Robin (2010) obtain non-parametric estimates of the distribution of hidden factors by performing three integrations (twice integrating the second derivative of the characteristic function of the factors, and once more to find the inverse Fourier transformation of the characteristic function). Their assumptions of additivity and independence of factors grant them analytic formulae and imply that all integrals to be computed are univariate. The counterpart of the latent factors in Bonhomme and Robin would be our VAR parameters, but since it is key to incorporate the covariances of the parameters (see the example in section 2) we would have to integrate *jointly* over hundreds of VAR parameters, hence a direct application of Bonhomme and Robin's approach would be numerically unfeasible.

Carrasco and Florens (2011) also estimate non-parametrically the probability distribution function of a hidden variable. The algorithms they propose involve solving large non-linear systems of equations. Available algorithms of the Gauss-Newton type involve inverting a matrix at each iteration, and this would be unfeasible in the very high-dimensional problem we consider. Our algorithm avoids any matrix inversion.

posed problems.[6] More work on the relationship between regularization and the approximate conjugate algorithm would be useful.

Many available algorithms for solving inverse problems need to restrict the probabilities to be non-negative and to add up to 1 at each step. These restrictions involve additional numerical complications. Another advantage of our algorithm is that it obtains proper densities at each step of the algorithm by construction.

Related to our work is the algorithm of Newton (2002) iterating on Bayes' formula. This algorithm is receiving recent attention in the non-parametric estimation literature. It is an on-line estimator (also called 'recursive' estimator in statistics), i.e., each observation is added one by one without updating previous estimates. On-line estimation is useful when relevant information arrives very rapidly, faster than the new information can be processed optimally by a computer.[7] It has also been a useful tool to obtain convergence results in the literature of least squares learning.[8] But these estimators add noise and inaccuracies in the estimation, so they are less justified in research papers. For example, one well-known side-effect of on-line estimation is that Newton's estimates depend on the ordering of the observations, it is also well known that that they are less efficient estimators. In ongoing research we investigate the application of our algorithm (described in section 3) to non-parametric estimation and we compare its properties to Newton's algorithm using our Proposition 5. Preliminary results indicate that Newton's algorithm is a noisy version of our

[6]For example, Carrasco and Florens (2011) use a Tikhonov regularization for the same purpose.

[7]Think of steering a ship into a harbor, where the angle of a rudder has to adjust to the direction of the wind; or think of choosing an optimal portfolio in a very unstable financial market. In such applications updating quickly the current value of the estimated quantity in view of a sudden change in the wind or on stock prices is likely to be more important than, say, maximizing the likelihood function using all past information as each new piece of information arrives.

[8]See Marcet and Sargent (1989) and Evans and Honkapohja (2002).

algorithm, that it converges much more slowly as the sample grows and that it has certain convergence problems which can be corrected by our approximate algorithm.

2 Priors about observables

Consider a model summarized in the likelihood function $p_{Y|\theta}$ that relates the distribution of the observable data Y to unknown parameters θ. Standard Bayesian practice is to find the posterior of θ after first stating a subjective prior p_θ directly on parameters. But for reasons discussed in the introduction it is desirable to use prior information about the observable data Y instead and to specify a prior on observables p_Y. The uncertainty represented in this prior can be seen as a combination of the researcher's uncertainty about the values of parameters θ and the error terms of the model $p_{Y|\theta}$. To find the posterior that incorporates the prior information contained in p_Y we first translate this prior on observables into a prior on parameters p_θ that is consistent with the model at hand. Then one can apply Bayes' formula in a standard way to obtain the posterior that is consistent with p_Y.

To demonstrate how a prior on observables can be translated into a prior on parameters we now use a simple example. This example will also serve to discuss issues of uniqueness and existence.

2.1 An example

Let variable y follow a univariate AR(1) model

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t, \text{ with } \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \text{ i.i.d.}, t = 1, ..., T. \tag{1}$$

$\mathcal{N}$ denotes the normal density. We treat y_0 and σ_ε^2 as given.

Most researchers would have a prior idea about the behavior of y in certain periods. In particular, one may have an idea about the distribution of y_1 and express this idea

by formulating a prior on the growth rate of y in the first period, for example,

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

for given $\mu_\Delta, \sigma_\Delta^2$. This distribution is compatible with many values of ρ and in no way it is saying that y follows a unit root. Although we do not write it explicitly to conserve space, the prior is conditional on the starting point y_0, hence (2) amounts to a prior on the behavior of y_1.

From the Bayesian point of view it is important that this prior should not be based on the estimation sample. It should come from other samples or other considerations. This requirement is the same as in the case of a standard prior on parameters.

For convenience, in this simple example we assume the prior in (2) is normally distributed, a known and fixed σ_ε^2, and we state the prior only about the first observation $t = 1$. The numerical methods we derive later in this paper do not need any of these features.

To translate the prior on observables (2) into the implied prior on α, ρ note that, given the AR(1) model

$$\begin{aligned}
\mu_\Delta &= E(\Delta y_1) = E(\alpha + (\rho - 1)y_0), \\
\sigma_\Delta^2 &= \mathrm{Var}(\Delta y_1) = \mathrm{Var}(\alpha + (\rho - 1)y_0) + \sigma_\epsilon^2.
\end{aligned}$$

This shows that a flat prior on α implies $\sigma_\Delta^2 = \infty$ and it says that the analyst addresses data with the prior that y is very likely to grow by more than, say, 100% at $t = 1$. But, as discussed in the introduction, the growth rate of an observable variable is not some abstract parameter and in most cases the analyst should be able to specify a much tighter prior. One message from this paper is that Bayesian researchers should at least examine the implications that their prior on parameters has for the behavior of the observables.

Our proposal, however, is to use a prior on observables, such as (2) to derive

a posterior consistent with it. Provided that $\sigma_\Delta^2 \geq \sigma_\epsilon^2$ the implied prior on (α, ρ) satisfies:

$$\alpha + (\rho - 1)y_0 \sim \mathcal{N}(\mu_\Delta, \sigma_\Delta^2 - \sigma_\epsilon^2). \tag{3}$$

This example brings about three points. First, for an arbitrary prior on observables there *may not exist* an implied prior on parameters that is compatible with the model, this would be the case if we had specified a prior variance on observables $\sigma_\Delta^2 < \sigma_\epsilon^2$. Second, there may be more than one solution, since (3) only imposes a restriction on a linear combination of α, ρ. To obtain a proper prior on parameters we need to complement (3) with an additional assumption, for example, about the marginal distribution of α or about the distribution of Δy_2. Third, equation (3) and the distribution of α imply a joint distribution of α and ρ with some non-zero correlation between α and ρ. This shows that the key in translating a prior on observables is to find the *joint* distribution of parameters. Many VAR applications assume priors in which parameters are mutually independent, this is understandable because specifying prior correlations between parameters is difficult, but imposing zero prior correlation on parameters often leads to unreasonable priors on observables. As we see in (3) a prior on observables is a natural way to specify such correlations among parameters.

2.2 A formulation as an inverse problem

We now return to the general case. Let Y take values on the space $\mathcal{Y}$ and θ take values on the space Θ. A key condition relating the prior on observables p_Y and the prior on parameters p_θ is

$$\int_\Theta p_{Y|\theta}(\overline{Y}; \cdot)\ p_\theta = p_Y(\overline{Y}) \quad \text{for almost all } \overline{Y} \in \mathcal{Y} \tag{4}$$

where the 'almost all' statement is with respect to p_Y. Note that p_Y is known given our stated prior on observables, and the likelihood function $p_{Y|\theta}$ is also known after

specifying a model. Our task is, given p_Y and $p_{Y|\theta}$, to find the prior density p_θ that satisfies the functional equation (4). This is known in calculus as 'a Fredholm equation of the first kind' and in statistics as an 'inverse problem' or identification of mixtures.

In the theoretical analysis we will assume that a solution p_θ exists, in practice we can ensure this in several ways by adjusting p_Y. Multiple solutions might arise, for example when the dimension of θ is larger than the dimension of Y, as in the AR(1) example above. See the empirical application in section 4.2 for one approach to selecting one from the potentially multiple solutions.

3 Fixed point formulation

Fredholm equations such as (4) can rarely be solved analytically.[9] Furthermore, they are not easy to solve numerically. For example, it may seem that an approximation can be easily found by discretising Y,θ and inverting a matrix version of the likelihood $p_{Y|\theta}$ at the discretised values of Y,θ to obtain a discretised approximation to p_θ. However, it is well known that the matrix's inverse for this case is ill-conditioned, the solution is unstable, and it often leads to a solution where p_θ has negative values and therefore is not a probability vector.

We now reformulate our inverse problem in terms of a fixed point problem that facilitates computation. We show conditions guaranteeing that this fixed point is

[9]The AR(1) example of section 2.1 is an exception. An analytic solution is available in that case because the growth rate of y in period $t = 1$ is linear in the parameters and both the prior on observables and the error ε are Gaussian. But with minor changes the analytic solution is no longer available. For example, if we would state a prior on the growth rates in *two* periods, $t = 1, 2$, then parameters enter non-linearly and an analytic solution is no longer available. The change of variable formula does not help either, see the online Appendix G for a further discussion.

necessary and sufficient for a solution to (4). We then propose an algorithm to compute this fixed point by successive approximations and we prove that this algorithm converges for the case when Y is continuous and θ is discrete. We finally describe approximate conjugate fixed point iterations that we use in practice to speed up computation and we show how to check for accuracy.

Let $g : \Theta \to \mathcal{R}_+$ be a probability density on Θ, in other words, g is a possible prior on parameters. Define the mapping $\mathcal{F}$:

$$\mathcal{F}(g)(\overline{\theta}) \equiv \int_{\mathcal{Y}} \frac{p_{Y|\theta}(\overline{Y};\overline{\theta})\ g(\overline{\theta})}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g}\ p_Y(\overline{Y})\ d\overline{Y} \quad \text{for all } \overline{\theta} \in \Theta. \tag{5}$$

The mapping $\mathcal{F}$ is indexed by $p_{Y|\theta}$ and p_Y but we leave this dependence implicit to avoid notational clutter.

$\mathcal{F}(g)$ has the following interpretation: let $p^g(\overline{\theta}|\overline{Y}) \equiv \frac{p_{Y|\theta}(\overline{Y};\overline{\theta})\ g(\overline{\theta})}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g}$, clearly p^g is the posterior distribution of θ obtained with the prior distribution g and given data realization $\overline{Y}$ is observed. Therefore, $\mathcal{F}(g)$ is the marginal density of θ when the joint distribution of (Y, θ) is given by $p^g p_Y$. Clearly, $\mathcal{F}(g)$ is a density so that $\mathcal{F}(g) \geq 0$ and $\int \mathcal{F}(g) = 1$.

We now show that there is a close relation between solutions to (4) and fixed points of $\mathcal{F}$.

Proposition 1. ***(Necessity)*** *If p_θ satisfies (4) then p_θ is a fixed point of $\mathcal{F}$.*

Even though necessity does not need any additional assumption, the following completeness condition is needed to establish sufficiency.

Definition 1. *Consider two random vectors a and b, each taking values in $\mathcal{A}$ and $\mathcal{B}$. Their joint distribution $p_{a,b}$ is said to be "complete with respect to a" when it holds that if a measurable function $\delta : \mathcal{A} \to R$ satisfies $E(\delta(a) \mid b) = 0$ for almost all $b \in \mathcal{B}$ then $\delta = 0$ a.s. in $\mathcal{A}$.*

We apply this definition to the joint distribution of Y, θ. Completeness with respect to θ is otherwise known as 'strong identification,' for example in Florens et al. (1990). The relationship between completeness and identification is a delicate issue. In the discrete θ case described in section 3.1.1 below the two notions are equivalent, but if θ is continuous completeness is stronger than identification.[10]

Proposition 2. ***(Uniqueness)*** *Assume there exists a solution to (4) satisfying $p_\theta > 0$. If $p_{\theta,Y}$ is complete with respect to θ, then p_θ is the unique solution to (4).*

(For a similar proposition see Florens et al. (1990), Theorem 5.5.20.)

Proposition 3. ***(Sufficiency)*** *Assume that $p_{\theta,Y}$ is complete with respect to Y. Then any fixed point $g^* = \mathcal{F}(g^*)$ such that $g^* > 0$ gives a solution to (4).*

The above propositions suggest that we can search for candidate solutions to (4) by finding fixed points of the mapping $\mathcal{F}$. If all the completeness and non-negativity conditions are satisfied this is guaranteed to deliver the unique solution to (4). Determining sufficient conditions for completeness in the continuous case is of interest but beyond the scope of this paper.[11]

Note that if the completeness conditions were to fail, this only affects the sufficiency and uniqueness propositions. We can guard ourselves from a failure of sufficiency in a given application by checking accuracy of a converged fixed point using the approach described in section 3.3. To explore possible nonuniqueness one should run the algorithm many times from different starting points.

[10] Furthermore, identification given a sample is weaker than identification of a mixture.

[11] There are some negative results in the literature, showing that completeness may be difficult to establish in non-parametric estimation setups, see Newey and Powell (2003); Canay et al. (2013). This should be less of an issue in our case since VAR likelihoods are highly parametric.

3.1 Succesive approximations on $\mathcal{F}$

Let us state a simple algorithm to search for fixed points of $\mathcal{F}$ by successive approximations. Let z denote the iteration number, we then define the following

Algorithm 1. ***(Successive approximations on $\mathcal{F}$)*** *1) Consider g^0, an initial density of θ. 2) Given g^{z-1} set $g^z = \mathcal{F}(g^{z-1})$. Repeat 2) for $z = 1, 2, ...$ until convergence.*

Algorithm 1 avoids the difficulties in solving inverse problems that we described at the beginning of this section: inversion of large matrices is entirely avoided and g^z is guaranteed to be a proper density at every iteration z.

As stated in Proposition 3 we need to ensure $g^* > 0$ to guarantee that the fixed point gives the desired solution. It is possible to see that $\mathcal{F}$ has 'false fixed points': there may exist $g^{**} = \mathcal{F}(g^{**})$ where $g^{**} = 0$ for some θ that do not satisfy (4). This serves as a word of caution: a good algorithm will stay away from densities that can be zero in some range of θ.

3.1.1 Convergence of successive approximations on $\mathcal{F}$

We now provide analytical results on the convergence of the successive approximations on $\mathcal{F}$. This proof is challenging, as it does not rely on standard techniques for convergence in economics, so we prove the result for the special case where θ takes on discrete values and we leave the continuous case for further research.

In the *discrete case* θ can only take N values $\{\theta_1, ..., \theta_N\} = \Theta$. The inverse problem now is to find an $N-$dimensional probability vector $p_\theta = [p_{\theta,i}]_{i=1}^N$ satisfying

$$\sum_{i=1}^{N} p_{Y|\theta}(\overline{Y}; \theta_i)\ p_{\theta,i} = p_Y(\overline{Y}) \quad \text{for almost all } \overline{Y} \in \mathcal{Y}. \tag{6}$$

(analogous to the continuous case where the problem was to find a density p_θ satisfying

(4)). The $\mathcal{F}$-mapping in the discrete case is

$$\mathcal{F}(g)_i \equiv \int_{\mathcal{Y}} \frac{p_{Y|\theta}(\overline{Y};\theta_i)\ g_i}{\sum_k p_{Y|\theta}(\overline{Y};\theta_k)\ g_k}\ p_Y(\overline{Y})\ d\overline{Y} \quad \text{for } i = 1, ..., N. \tag{7}$$

Note that we abuse notation slightly since in the continuous case p_θ, g and $\mathcal{F}(g)$ were densities while these symbols denote $N-$dimensional vectors in this subsection. A further abuse of notation is that $p_{\theta,Y}$ denotes a joint probability while $p_{\theta,i}$ is the $i-$th element of p_θ.

It is easy to adapt the proofs of Propositions 1 to 3 to show that the results also hold for the discrete case. The case of continuous Y and discrete θ has a long tradition in the statistics and probability literature, hence much is known about completeness in this case. For example, Teicher (1963) Theorem 1 shows that identifiability of a finite mixture is equivalent to completeness and Teicher (1963) Proposition 1 that a finite mixture of normals is identified.

The following theorem supports Algorithm 1.

Proposition 4. ***(Convergence)*** *Assume that i) $p_{\theta,Y}$ is complete with respect to θ, ii) θ is discrete, $\{\theta_1, ..., \theta_N\} = \Theta$, iii) there is a solution to (6) with $p_{\theta,i} > 0$ for all i.*

Then all eigenvalues of the derivative $\frac{\partial \mathcal{F}(p_\theta)}{\partial g'}$ are real and they belong to the interval $[0, 1)$.

Therefore, successive approximations on $\mathcal{F}$ converge locally to p_θ.

Formally, this means that letting g^z be the vector defined in Algorithm 1, there is an open neighborhood $S \subset \left\{g \in R^N_{++} : \sum_i g_i = 1\right\}$, of $p_\theta \in S$, such that for all $g^0 \in S$ we have $g^z \to p_\theta$ as $z \to \infty$.

3.2 Approximate conjugate algorithm

We now propose a practical numerical algorithm based on *approximate* iterations on the mapping $\mathcal{F}$ when Y and θ are general continuous random variables. This

approximate conjugate algorithm is the one we apply to real life applications in section 4. At each iteration we restrict the density g to be in a given parametric family that is conjugate with the likelihood. Conjugacy speeds up the iterations and, later, the computation of the posterior. We place no restriction on the density p_Y except that it must be possible to generate draws from this distribution on a computer.

Of course, fixing a parametric family is a good approach only as long as the solution of the inverse equation (4) is approximated with the desired accuracy by the proposed parametric family. We discuss how to check ex-post if the accuracy of the approximation is acceptable in section 3.3.

Let $\mathcal{G}$ be a given parametric family of densities on Θ. Let $q : \Theta \to R^\nu$ be a function such that the moments $E_p(q(\theta))$ suffice to pin down any density $g \in \mathcal{G}$.[12]

Algorithm 2. ***(Parameterized successive approximations on $\mathcal{F}$)***

1) Start with an initial density $g^0 \in \mathcal{G}$

2) Given $g^{z-1} \in \mathcal{G}$ compute the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$.

3) Let $g^z \in \mathcal{G}$ be given by the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$.

Repeat 2)-3) until convergence of the moments $E_{\mathcal{F}(g^z)}(q(\theta))$.

In words, we obtain each successive iteration $g^z \in \mathcal{G}$ by projecting $\mathcal{F}(g^{z-1})$ back onto the family $\mathcal{G}$. Typically, the moments involved in Step 2 will need to be approximated numerically. When $\mathcal{G}$ is conjugate one can approximate these moments efficiently using the following result. Recall the definition of p^g after equation (5), then

Result 1. *Given any density g, for any function $q : \Theta \to R^\nu$ we have*

$$E_{\mathcal{F}(g)}(q(\theta)) = E_{p_Y}\left[E_{p^g(\cdot|Y)}(q(\theta))\right]. \tag{8}$$

[12]For example, $\mathcal{G}$ can be the set of Gaussian densities. In that case $q(\theta) \equiv (\text{vec}(\theta), \text{vec}(\theta\theta'))$.

This result can be used to speed up the computation of the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$ required in Step 2 by exploiting analytic conjugate expressions for moments $E_{p^g(\cdot|Y)}(q(\theta))$ as follows: draw J realizations of Y from p_Y, then split Step 2 into two steps: 2a) For each realization $\overline{Y}$ compute (if possible, analytically) the posterior moments of θ using g^{z-1} as the prior, that is $E_{p^{g^{z-1}}(\cdot|\overline{Y})}(q(\theta))$. 2b) Approximate $E_{p_Y}[\cdot]$ in (8) by averaging the posterior moments obtained in Step 2a over the J draws of Y. If the family of conjugate priors $\mathcal{G}$ is such that the moments in Step 2a) are available in closed form this computation can be done very efficiently. When $\mathcal{G}$ is not conjugate then Algorithm 2 is slower because a separate Monte Carlo procedure is needed for each draw $\overline{Y}$ in order to evaluate the moments $E_{p^g(\cdot|\overline{Y})}(q(\theta))$.

As a simple example of the above procedure we now write in detail a special case of this algorithm for the example in section 2.1, where $\theta = (\alpha, \rho)$, the likelihood $p_{Y|\theta}$ is given by the model specified in (1), with a known σ_ε^2, and supposing that $\mathcal{G}$ is the class of normal distributions. The normal distribution is conjugate in this model. Consider a prior on observables p_Y describing the behavior of (y_1, y_2), hence an analytic solution is not available (see footnote 9). Let $M^{pri} \equiv E_{pri}(\theta)$ and $\mathcal{V}^{pri} \equiv E_{pri}(\theta\theta')$ be the prior (and $M^{po}, \mathcal{V}^{po}$ the posterior) mean and second moment of θ. Given a sample $\overline{Y} = (\overline{y}_1, \overline{y}_2)$ a standard result in Bayesian statistics fully characterizes the posterior as given by the moments

$$(M^{po}, \mathcal{V}^{po}) = F_{\mathcal{N}}(M^{pri}, \mathcal{V}^{pri}; \overline{Y}) \tag{9}$$

for a well known function $F_{\mathcal{N}}$. Then we can combine Algorithm 2 with Result 1 in the following

Algorithm 2.A. *$\mathcal{G}$ be the class of normal distributions.*

Draw J independent realizations $\overline{Y}^j$ from p_Y, J a large integer.

1) Start with an initial $g^0 \in \mathcal{G}$ with mean $M^0 = E_{g^0}(\theta)$ and second moment $\mathcal{V}^0 = E_{g^0}(\theta\theta')$.

2) Given a prior $g^{z-1} \in \mathcal{G}$ with mean M^{z-1} and second moment $\mathcal{V}^{z-1}$ approximate $E_{\mathcal{F}(g^{z-1})}(\theta, \theta\theta')$ with $(M^z, \mathcal{V}^z) = \frac{1}{J}\sum_{j=1}^{J} F_{\mathcal{N}}(M^{z-1}, \mathcal{V}^{z-1}; \overline{Y}^j)$.

3) Set the next iteration $g^z \in \mathcal{G}$ with mean and second moment $M^z, \mathcal{V}^z$.

Repeat 2)-3) until convergence of M^z and $\mathcal{V}^z$.[13]

The result is a normal approximate fixed point of $\mathcal{F}$.

Algorithm 2.A shows how Algorithm 2 and Result 1 can be combined in a simple case.[14] But Algorithm 2.A assumes that the innovation variance σ_ε^2 is known. In most practical applications this variance is not known. In the next algorithm we incorporate uncertainty about the innovation variance and generalize to the case of a multivariate VAR. We set $\mathcal{G}$ as the family of Normal-Inverted Wishart conjugate prior densities of the parameters of a VAR model and combine Algorithm 2 with Result 1. Here is a full description of this algorithm that we use in the empirical applications in section 4.

The VAR model for the $N \times 1$ vector of observables y_t is

$$y_t = \sum_{p=1}^{P} B_p \; y_{t-p} + c + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma), \quad t = 1, ..., T. \tag{10}$$

The parameters are $\theta = (B, \Sigma)$, for a matrix $B = (B_1, ..., B_P, c)'$, P is the number of lags, the initial values $y_{-P+1}, ..., y_0$ are treated as fixed and the analysis conditions on

[13]Usually normal distributions are expressed in terms of variances V instead of second moments $\mathcal{V}$. Obviously either choice is equivalent taking $V = \mathcal{V} - MM'$. We use second moments in the main text because then the formulae in Algorithm 2.A are simpler. Had we used variances we would have to use in step 2 the longer, but equivalent expression $V^z = \frac{1}{J}\sum_j^J F_V(M^{z-1}, V^{z-1}; \overline{Y}^j) + \frac{1}{J}\sum_j^J F_M(M^{z-1}, V^{z-1}; \overline{Y}^j) F_M(M^{z-1}, V^{z-1}; \overline{Y}^j)' - M^z M^{z\prime}$ for well known functions $F_M(M^{pri}, V^{pri}; \overline{Y})$ and $F_V(M^{pri}, V^{pri}; \overline{Y})$ that give the posterior mean and variance in a linear Gaussian model.

[14]For an application see Jarociński and Lenza (2018, section B.4).

them. The Normal-Inverted Wishart conjugate prior density of B and Σ satisfies

$$p(\operatorname{vec} B|\Sigma) = \mathcal{N}(\operatorname{vec} M, \Sigma \otimes Q), \tag{11}$$

$$p(\Sigma) = \mathcal{IW}(S, v), \tag{12}$$

where $\mathcal{IW}$ denotes the Inverted Wishart density and M, Q, S, v are prior parameters of appropriate dimensions.

As in Algorithm 2.A we denote $M = E(B)$ and $\mathcal{V} = E(\operatorname{vec} B(\operatorname{vec} B)')$. We also denote the moments of Σ^{-1} as $D = E(\Sigma^{-1})$ and $\mathcal{H} = \operatorname{diag} E(\operatorname{vec} \Sigma^{-1} (\operatorname{vec} \Sigma^{-1})')$. Analogous to (9), given a Normal-Inverted Wishart prior with parameters $(M^{pri}, Q^{pri}, S^{pri}, v^{pri})$ and a sample $\overline{Y}$, the posterior moments are given as

$$(M^{po}, \mathcal{V}^{po}, D^{po}, \mathcal{H}^{po}) = F_{\mathcal{NIW}}(M^{pri}, Q^{pri}, S^{pri}, v^{pri}; \overline{Y}) \tag{13}$$

for a well known function $F_{\mathcal{NIW}}$. For completeness we derive closed form expression for $F_{\mathcal{NIW}}(M^{pri}, Q^{pri}, S^{pri}, v^{pri}; \overline{Y})$ in the Online Appendix. Then we can use

Algorithm 2.B. *$\mathcal{G}$ be the class of Normal-Inverted Wishart distributions.*

Draw J independent realizations $\overline{Y}^j$ from p_Y, J a large integer.

1) Start with an initial prior $g^0 \in \mathcal{G}$ given by parameters M^0, Q^0, S^0, v^0.

2) Given $g^{z-1} \in \mathcal{G}$ with parameters $M^{z-1}, Q^{z-1}, S^{z-1}, v^{z-1}$, approximate the relevant moments given the density $\mathcal{F}(g^{z-1})$ with

$$(M^z, \mathcal{V}^z, D^z, \mathcal{H}^z) = \frac{1}{J} \sum_{j=1}^{J} F_{\mathcal{NIW}}(M^{z-1}, Q^{z-1}, S^{z-1}, v^{z-1}; \overline{Y}^j)$$

3) Find parameters M^z, Q^z, S^z, v^z so as to match the moments $M^z, \mathcal{V}^z, D^z, \mathcal{H}^z$ as closely as possible with a Normal-Inverted Wishart density. Let $g^z \in \mathcal{G}$ be given by parameters M^z, Q^z, S^z, v^z.

Repeat 2)-3) until convergence of M^z, Q^z, S^z, v^z.

One difference with Algorithm 2.A is that step 3) is no longer automatic, because the Normal-Inverted Wishart density is not parameterized directly in terms of its moments. In fact, the Normal-Inverted Wishart density imposes certain constraints on the first two moments, so in general one cannot match the moments $M^z, \mathcal{V}^z, D^z, \mathcal{H}^z$ exactly. The approach we follow in practice is to match M^z and D^z exactly, and to match $\mathcal{V}^z$ and $\mathcal{H}^z$ approximately using closed form expressions for M^z, Q^z, S^z, v^z that we show in the Online Appendix.

3.3 Accuracy checking

After performing the iterations the algorithm will have reached a solution, say, g^Z. It is clear that g^Z will not satisfy (4) exactly. First because the iterations might not reach an exact fixed point of $\mathcal{F}$. Second because we use an approximate conjugate algorithm as described in the previous subsection. Third because in practice it is difficult to know if an exact solution to (4) exists. Therefore we need to check for accuracy.

Letting $p_Y^Z = \int_\Theta p_{Y|\theta}\ g^Z$, we check accuracy by comparing p_Y^Z and p_Y, if they were exactly equal g^Z would be the solution we seek. Furthermore, g^Z would be the solution we seek if the prior on observables were p_Y^Z in the right side of (4).Therefore, when a solution to (4) does not exist but p_Y^Z is 'reasonably close' to p_Y then g^Z should be an acceptable translation of p_Y, after all the prior densities p_Y that a researcher may state for observables can only be indicative.

For this purpose we compute moments or interval frequencies from a large number of draws of p_Y^Z and p_Y. Draws from p_Y^Z are straightforward to obtain as follows: draw a realization of parameter values $\overline{\theta}$ from the approximate fixed point g^Z, and then draw Y from $p(\cdot|\overline{\theta})$. We apply this procedure in our empirical applications below. For example, as an advance of future results, the reader can now glance at Figure

3 plotting the quantiles of the prior on observables p_Y (blue shaded area) and the quantiles of the distribution of the observables implied by the approximate fixed point p_Y^Z (solid line). As can be seen these are very close.

Also, as an example, we do a Monte Carlo experiment to study the performance of the approximate fixed point algorithm. We use a setup where problem (4) has a known high-dimensional solution p_θ and check if our algorithm recovers this solution. With random starting points g^0 the algorithm always recovers the 667 parameters that index p_θ with great precision in under 5 minutes on a standard personal computer. Details of this Monte Carlo experiment are in the Online Appendix.

4 Empirical Applications

This section presents two applications of priors on observables to the estimation of structural VARs. Both examples are well known VARs that have been estimated many times in the literature. Example 1 is the fiscal policy VAR of Blanchard and Perotti (2002). Example 2 is the study of the effects of monetary policy shocks by Christiano et al. (1999).

The aim of this section is to make five points. First, different standard priors on parameters available in the literature can give significantly different results and, as VAR parameters are hard to interpret, there are few reasons a-priori to choose among these alternatives. Second, some of these priors on parameters imply priors on observables that are unlikely to represent the prior knowledge of the analyst. Third, the algorithm proposed in section 3.2 is feasible in practical applications and that it gives an accurate solution to the inverse problem. Fourth, the examples show how to set up the prior on observables in various ways: in the first example the prior summarizes the ideas expressed by the authors of the original paper about the likely behavior of the variables, while in the second example we use an empirical

Bayes prior. Fifth, in these applications the prior on observables affects the results, in fact changing them considerably. To the extent that this prior on observables is a better representation of the analyst's prior knowledge, we contend that the resulting posterior is better justified from the subjective Bayesian point of view.

We use four standard priors for θ as the reference. The first one is the flat (non-informative) prior, where the posterior mean of B is the OLS estimate. Both papers from which we take our examples, Blanchard and Perotti (2002) and Christiano et al. (1999), use the OLS estimation, hence the flat prior comes closest to replicating their results (apart from small discrepancies between their bootstrap and our Bayesian uncertainty bands). The remaining three are standard informative priors for VARs in the Litterman, Sims and Zha tradition, using three off-the-shelf choices for the hyperparameters. We refer to them as the 'Minnesota' prior (the default prior in the RATS computer package), the 'Sims Zha (1998)' prior (a widely used version of the prior) and the 'Dynare' prior (the default prior in the Dynare computer package). See Appendix B for the precise definitions of these priors.

We then apply our approach. In each application we use a simple auxiliary model to construct the prior density of observables. The auxiliary model in each case is such that all its parameters have a clear interpretation in terms of the behaviour of observables, but the model is too simple to be of interest per se. Having specified a prior on observables we then translate this density into a prior for the VAR parameters using the algorithm described in section 3.2. Finally, we use Bayes' theorem in the standard way to compute the posterior.

The VARs are specified in levels and the variables entering them are clearly non-stationary. Therefore, the prior density of these variables must be conditional on some initial state. A natural choice is to use as the initial state the P first observations in the sample, where P is the number of lags. The VAR likelihood function

conditions on the same P observations, so it is logically consistent that the prior and the likelihood condition on the same initial state.

The structure of the presentation is the same in each example: we present the empirical application, show the results obtained with the standard priors for VARs, compare the implied prior on observables that emanates from these standard priors, state our prior about observables, study the accuracy of the algorithm in computing the translated prior, and finally we show the posterior implied by the prior about observables.

4.1 Blanchard and Perotti (2002) VAR

In this subsection we estimate the effects of tax and government spending shocks following Blanchard and Perotti (2002). Their VAR includes taxes, government spending and GDP, all in real, per capita terms, and the estimation sample is 1960Q1-1997Q4.[15] They identify structural shocks to taxes and spending using restrictions on the relations between reduced form residuals and structural shocks. Their key identifying restriction separates tax shocks from their endogenous responses using the elasticity of tax innovations to output innovations estimated separately from disaggregated data. Using priors about observables in this application is natural, as Blanchard and Perotti themselves state their beliefs about the relation between output, tax revenues and spending, beliefs that inspire our subjective prior on observables.

4.1.1 Results with standard priors

Figures 1 and 2 show the effects of, respectively, tax and spending shocks. We report quantiles 0.16 and 0.84 of the posterior distributions of the impulse responses (Blanchard and Perotti report one-standard-deviation bootstrap bands). The variables are

[15] We downloaded the data from Olivier Blanchard's webpage.

quarterly, in log levels, and we rescale the responses so that they correspond to a one percent shock to, respectively, taxes or spending. The blue shaded regions (common to all plots in a given row) report the posteriors obtained with the flat prior, so they are the closest to the OLS estimation of the VAR by Blanchard and Perotti.[16] The black lines report the posteriors obtained with informative priors, each column of graphs representing the results with a different estimation procedure. The first three columns are for the standard informative priors: the Minnesota prior, the Sims and Zha (1998) prior and the Dynare prior, and we ask the reader to disregard the fourth column for now.

Figure 1 shows that responses to a tax shock differ widely across standard VAR priors. What is common is that after a one percent tax shock taxes increase, spending falls with some delay, and GDP starts falling immediately, but the time profiles of these responses differ strongly. For example, under the flat prior taxes revert to the baseline after about 10 quarters, and under the Minnesota prior they are only marginally more persistent. By contrast, under the Sims and Zha (1998) prior taxes remain permanently higher by about 0.7 percent, while under the Dynare prior taxes remain permanently higher by about 0.45 percent. There are also differences in the responses of GDP: under the flat and Minnesota priors GDP falls, reaching -0.2 percent after about 10 quarters and then starts to gradually return to the baseline. Under the Sims and Zha (1998) and Dynare priors GDP falls by only about 0.07 and

[16]Blanchard and Perotti (2002) estimation has some nonstandard features: they estimate four sets of VAR coefficients, one for each quarter of the year, to account for seasonal patterns and they subtract time-varying stochastic trends or linear trends. Subsequent literature has followed Blanchard and Perotti's identification but mostly ignored these nonstandard features. We follow this literature and estimate standard VARs in levels (which is the closest to their specification with stochastic trends). Nevertheless, the impulse responses we obtain with approximately flat priors are similar to Blanchard and Perotti's impulse responses.

0.15 percent respectively, but this fall appears to be permanent. Figure 2 reports similarly large differences in responses to the spending shock, the most striking ones in the response of spending itself.

[Figure 1 about here.]

[Figure 2 about here.]

This shows that Bayesian VARs can produce very different results under different priors. These differences are relevant for evaluating the effects of austerity: the output costs of increasing taxes more than doubles with a flat prior compared with the Sims and Zha (1998) prior, and they are even larger for the Minnesota prior. The output costs of cutting spending are very uncertain with all the estimation procedures and for a given initial cut in spending they are the largest under the Dynare prior. However, most researchers will find little reason to choose one prior over another based on a priori grounds, because it is difficult to interpret priors on VAR parameters directly.

Furthermore, these priors on parameters imply priors about data behavior that no analyst would ever hold, hence they can not represent an analyst's prior information. Figure 3 reports the densities of each variable implied by plugging in the left side of (4) the corresponding prior on parameters. Thus the figure shows the prior on observables that would be consistent with the priors on parameters found in the literature. The figure plots the quantiles of the density of each variable for periods $t = 1, 2, ..., 15$ at the start of the sample. The blue shaded region shows the prior about observables expressed by Blanchard and Perotti in their paper, we describe this density in more detail below in section 4.1.2. This blue shaded region shows that uncertainty gradually increases as time goes by, as more error terms accumulate, consistent with the model used. It also shows that output is on average expected to grow. The solid black line gives the quantiles for the fixed point that we compute,

please ignore this line for now. The dashed and dotted lines show the quantiles for the priors on observables implied by the standard VAR priors used in estimation. We can see that in some cases these priors are quite counterintuitive. The Minnesota and noninformative (flat) priors are the most striking, as they place almost a uniform distribution on growth rates over the real line (the quantiles look vertical given the scale of the plot).[17] These priors imply, for example, that a yearly output growth of more than 100% is much more likely than a growth rate of between 0 and 4% a year. We contend that no analyst will deem this to properly represent his/her views about the economy. The other two priors are less unreasonable but still have some problems: Sims-Zha is centered on the scenario of zero output growth and Dynare on negative growth, while placing nonnegligible probability on very large positive or negative growth rates of some variables, e.g. taxes.

[Figure 3 about here.]

This figure is meant to show that the standard priors on parameters are unlikely to represent the opinion of the analyst in this application. Hence, the posteriors found with these prior distributions are not appealing on subjective Bayesian grounds. This is why we consider priors specified explicitly on observables instead.

4.1.2 A prior about observables

We now formulate a prior about observables, p_Y. The prior is about the dynamics of GDP, taxes and spending in the beginning of our estimation sample. We base this prior on the data from the period preceding the estimation sample, and on subjective priors about the relations of taxes and spending with GDP inspired by the comments in Blanchard and Perotti (2002). One aspect of these priors is cointegration, for

[17] This is a consequence of assuming a flat prior on the intercept so it should hold for any application of flat and Minnesota priors.

an alternative approach to priors about cointegration in VARs see Giannone et al. (2018).

The data that inform our prior are on real GDP for the period 1947-1960 and on taxes and spending in 1960.[18] We fit an AR(2) model into the GDP data for 1947Q1 to 1960Q4 and generate the predictive density of GDP after 1960Q4. Then, following Blanchard and Perotti (2002), we consider cointegration relations between variables, and we use their model of innovations. Specifically, we postulate that taxes and spending are cointegrated with GDP and follow

$$\tau_t = \delta_x + \tau_{t-1} - \beta^\tau(\tau_{t-1} - x_{t-1} - c^\tau) + a_1 u_t^x + \sigma^\tau \varepsilon_t^\tau, \tag{14a}$$

$$g_t = \delta_x + g_{t-1} - \beta^g(g_{t-1} - x_{t-1} - c^g) + \sigma^g \varepsilon_t^g, \tag{14b}$$

where τ_t is the log of taxes, g_t is the log of spending, u_t^x is the innovation to GDP, ε_t^τ and ε_t^g are the tax and spending shocks, both i.i.d. standard normal random variables, and δ_x, β^τ, β^g, c^τ, c^g, a_1, σ^τ, σ^g are scalar parameters. We set the constant term δ_x equal to the average growth rate of GDP in the 1947Q1-1960Q4 sample. We set c^τ, c^g, i.e. the logs of the equilibrium shares of taxes and spending in GDP, to the average values of $\tau_t - x_t$ and $g_t - x_t$ in 1960 (where x_t is the log of GDP). We set $\beta^\tau = \beta^g = 0.5$, implying a fast convergence of taxes and spending to these equilibrium shares in GDP. We assume that the standard deviations of tax and spending shocks, σ^τ and σ^g, are both 1%. Finally, $a_1 = 2.08$ is the elasticity of tax innovations to GDP innovations that Blanchard and Perotti estimated from disaggregated data and used

[18] Blanchard and Perotti's estimation starts in 1960Q1, but it is ok to use the data from 1960 to inform our prior because the VAR has four lags and when estimating it we condition on the data for the four quarters of 1960 anyway. The replication dataset does not include taxes before 1960Q1. Moreover, as discussed in Blanchard and Perotti (2002), government spending before 1960Q1, while available in the replication dataset, is unusually volatile due to the Korean War expenditures in the 1950s, so in our baseline prior about observables we ignore these data and use only GDP before 1960.

in their VAR identification. They argue that the elasticity of spending innovations to GDP innovations is zero, hence we do not include u_t^x in the equation for spending. The implied predictive density of taxes, spending and GDP is our prior about observables. We impose this predictive density for 15 quarters.[19] We have plotted draws from the above prior density and both their dynamics and comovement do resemble plots of actual GDP, taxes and spending.

After specifying this density of the observables we run the approximate conjugate algorithm from section 3.2 where $\mathcal{G}$ is the family of Normal-Inverted Wishart densities (see the Online Appendix for the details on the implementation). Using different random starting points g^0, the algorithm always converges to a similar Normal-Inverted Wishart density. In what follows we present results based on one thousand iterations on the algorithm, which take about 12 minutes on a standard PC, but we obtain very similar results already after 200 iterations.[20]

Finally, we check the accuracy of the fixed point that we find by comparing the implied density of observables with the stated density in the prior. Figure 3 shows with the shaded region the quantiles according to the prior on observables. The solid

[19]We follow an informal rule of thumb to specify the dimension of the prior about the observables equal to the dimension of the prior about parameters. E.g. here the dimension of the prior density of the observables (15×3) equals the dimension of the prior density of the parameters B and Σ (i.e. $N(NP+1)+N(N+1)/2 = 45$). In our experiments we find that when the prior dimension satisfies this rule of thumb our approximate conjugate algorithm converges to a unique fixed point, and when the prior dimension is much lower (as in section 4.2) there are multiple fixed points. Theory suggests that the inverse problem (4) may have a unique solution even when the dimension of p_Y is 1, so we stress the informal and empirical nature of this rule of thumb.

[20]Somewhat disappointingly, in this example (unlike in the next one) the marginal likelihood implied by our prior on observables is lower than those implied by the standard priors. In future research we would like to understand better what features of the priors on observables are needed for a higher marginal likelihood.

lines are the quantiles with our approximate fixed point. As can be seen the match is nearly perfect.

4.1.3 Results with the prior about observables

The rightmost columns of Figures 1 and 2 report the responses to tax and spending shocks implied by the subjective prior about observables. The responses to the tax shock (Figure 1) are closest to those obtained with the Minnesota prior. The main difference is that the immediate response of spending is negative (instead of being close to zero) and, consistently with this, the negative response of output is slightly stronger. The responses to the spending shock (Figure 2) obtained with the prior about observables imply a larger government spending multiplier than according to any of the other methods. The response of output to a 1% shock is about 0.3% after 12 quarters, compared with about 0.2% according to the OLS estimation and Minnesota prior. The response of output obtained with the Dynare prior might be even higher than 0.3% after 12 quarters, but it is associated with a much higher spending after the initial shock (about 1.5% above the benchmark after 12 quarters, as opposed to less than 1% when the prior about observables is used). Summing up, the subjective prior about observables yields plausible impulse responses, with the effects of tax shocks on output that are more negative than under the flat prior and much more negative than under the Sims Zha (1998) and Dynare priors, and with more positive effects of spending shocks on output than under alternative priors. From the point of view of our prior about observables, standard VAR priors underestimate fiscal multipliers.

4.2 Christiano et al. (1999) VAR

In this subsection we estimate the effects of monetary policy shocks following Christiano et al. (1999) (CEE). They estimate a VAR in levels with output (real GDP),

prices, commodity prices, federal funds rate, total reserves, nonborrowed reserves and money, using quarterly US data from 1965 to 1995.[21] The monetary policy shock is identified as the Choleski shock to the federal funds rate, with the above ordering of the variables.

4.2.1 Results with standard priors

Figure 4 shows the effect of monetary policy shocks on output. We report the quantiles 0.05 and 0.95 of the posterior distributions of the impulse response of GDP (CEE report 90% bootstrap bands). Responses of the remaining variables are reported in the Online Appendix. GDP is quarterly, in log levels, and the responses correspond to a one standard deviation shock. The shaded regions (common to all four plots) report the posterior obtained with the flat prior, so they are the closest to the OLS estimation of the VAR by the CEE.

Panels A to C illustrate that the persistence of output responses differs dramatically depending on the prior on parameters used. The flat prior (shaded) produces a short-lived effect (the shaded 90% posterior probability range contains zero after about 10 quarters). The Minnesota prior in panel A produces similar persistence as the flat prior but narrower error bands. The Sims Zha (1998) prior in panel B and the Dynare prior in panel C tend to produce permanent responses of output (and, in panel C, a quite high probability of an explosive response). The permanent responses in panels B and C are inconsistent with the long-run neutrality of money and thus they pose a challenge to most standard economic theories, which almost always imply long-run neutrality of money. Again, as in the Blanchard and Perotti (2002) example, we find that different standard priors produce different results, so it is important to think about whether or not the priors can represent the anlysts' prior information.

[21] We downloaded the data from Larry Christiano's webpage.

[Figure 4 about here.]

Figure 5, analogous to Figure 3, plots over time the quantiles of the observables implied by different standard priors and we find that they miss on some key aspects. The Minnesota and noninformative (flat) priors are the most extreme ones as they imply that huge growth rates are very likely. The Sims Zha (1998) and Dynare priors are consistent with a zero average inflation and no growth of money supply, reserves and GDP. To the extent that this does not represent the analysts' opinion on the behavior of observables we conclude that the posterior is not convincing on subjective Bayesian grounds.

[Figure 5 about here.]

4.2.2 A prior about observables

This time we formulate a minimalistic prior about observables. The prior is about the initial growth rates of all the variables. We call it minimalistic for two reasons. First, it conveys very simple ideas about the dynamics of the observables, namely, that the observables follow independent random walks, shadowing the idea behind the priors in the Litterman, Sims and Zha tradition. Second, we specify this prior for only a few periods, fewer than necessary to define the density of parameters uniquely. This is because given the simplicity of the prior we do not want to impose it too dogmatically.

Our prior on observables is a $P \times N$ dimensional density $p_{\Delta y_1,\dots,\Delta y_P|y^o_{-P+1},\dots y^o_0}$. Recall that $P = 4$ is the number of lags.[22] Specifying a prior on growth rates does not mean we impose a unit root, it is done only for convenience, obviously this prior

[22] Jarociński and Marcet (2010), section 2, draw parallels between this prior and some frequentist small sample estimators.

is equivalent with a certain density for the levels $p_{y_1,...,y_P|y^o_{-P+1},...y^o_0}$. The density could be drawn from the purely subjective prior opinion of the user, but here we take an empirical Bayes approach and use the growth rates observed in the data to inform our prior.[23] Therefore, our prior conveys the idea that the growth rates of the first P observations behave similarly as the rest of the sample. The way we implement this idea is the following: we estimate an auxiliary model $\Delta y_{n,t} = \alpha_n + \varepsilon_{n,t}$, $\varepsilon_{n,t} \sim \mathcal{N}(0, \sigma_n^2)$ for each variable $n = 1, ...N$ and use as $p_{y_1,...,y_P|y^o_{-P+1},...y^o_0}$ the density of the observables implied by the posteriors of α_n, σ_n^2. In the Online Appendix we report the growth rates observed in our sample and discuss other variants of the prior that use data from various subsamples and from the period *preceding the estimation sample*.

The blue region in Figure 5, shows the distribution of observables implied by the empirical Bayes prior on observables. As we can see, it differs from the standard informative priors because output, prices, reserves and M1 are expected to grow over time.

4.2.3 Results with the prior about observables

We already saw in the context of example in section 2.1 that if the prior on observables involves fewer variables than the number of parameters in the model, then the implied prior on parameters might be improper. We expect something similar to happen in this application since we use a prior on $NP = 28$ observables, much lower than the

[23] The empirical Bayes approach is controversial because it makes the prior dependent on the data. The advantages and disadvantages of this approach have been discussed at length in the literature, see Morris (1983) for a classical reference or Efron (2010) for a more recent reference. Our use of the empirical Bayes approach here follows Berger (1985, section 3.5.2) who suggests the data themselves as a possible source of information about the marginal density of the data.

number of parameters ($N(NP+1)$ in B and $N(N+1)/2$ in Σ, 231 in total).[24] One way to proceed could be to complement the prior on observables with additional priors to define a unique prior. This could be implemented by weighing the obtained fixed points with this complementary prior. We proceed in a slightly different way: we find as many priors consistent with the prior on observables as we can, and we select the one with the highest marginal likelihood and the the one with the highest entropy. These choices somehow represent two opposite criteria: the highest marginal likelihood is the prior that best fits the data actually observed,[25] while maximum entropy can be interpreted as imposing as little prior knowledge as possible.[26] Roughly speaking, we could expect most other priors to stay between these two extremes.[27] Finally, we restrict the marginal prior density of Σ to be an Inverted Wishart density centered at the standard errors of the univariate autoregressions estimated by ordinary least squares for each variable. This is the same marginal density of Σ as in the Minnesota, Sims Zha (1998) and Dynare priors (see Appendix B).

We implement this by computing 300 approximate fixed points that satisfy the

[24]In general it is not necessary for identification to have the dimension of the prior on the observables weakly greater than the dimension of the prior on the parameters. However, in practice we often find multiplicity of solutions when the dimension of the prior on observables is much lower, see also footnote 19.

[25]The marginal likelihood is $\int p(y^o|\cdot)p_\theta$, where y^o is the observed data.

[26]Entropy, defined as as $\int_\theta \log p(\theta)dp(\theta)$ measures the amount of information carried by a distribution. We obtained an analytical expression for the entropy of a Normal-Inverted Wishart density with the help of Proposition 3 of Gupta and Srivastava (2010).

[27]It also happens to be the case that, among the priors we find consistent with the prior on observables, the maximum-marginal-likelihood has one of the lowest entropies, and that the maximum-entropy fixed point has one of the lowest marginal likelihoods. Therefore the two priors can be interpreted as having the two extreme entropies (or marginal likelihoods). In both cases the marginal likelihoods are higher than those implied by the standard priors.

restriction on $p(\Sigma)$, each with a different random starting point g^0. We use the approximate conjugate algorithm from section 3.2. We stop at 300 because the lessons drawn are the same as those based on the first 200. Finding each fixed point requires about 200 iterations and takes about 5 minutes with Matlab on a standard personal computer. From these 300 fixed points we choose the one with the highest marginal likelihood and the one with the highest entropy.

To check accuracy we look at the implications for observables of the approximate fixed points that we find. The solid lines in Figure 5 show the quantiles implied by the left hand side of (4) at a representative approximate fixed point with the restriction on $p(\Sigma)$. The solid lines are close to the edges of the shaded regions that represent our desired prior about observables. This shows that, in spite of its approximate nature, its very large dimensionality and the restriction on $p(\Sigma)$, the approximate conjugate algorithm delivers a density of observables that is reasonable and close to the desired prior.[28]

The posterior for the fixed point with the highest marginal likelihood in the sample is plotted with the solid line in panel D of Figure 4. The posterior shows a much more persistent effect of monetary shocks than OLS: output takes about 20 quarters to recover, instead of about 10 quarters with the flat prior. The effect of the shock in the first two years is weaker with our prior but it becomes stronger afterwards. The median total output loss after 5 years is 30% larger according to our prior than with the flat prior (1.85% of yearly output loss in our case versus 1.40%).[29] More

[28]In the absence of the restriction on $p(\Sigma)$ we find fixed points for which the solid lines are indistinguishable from the edges of the shaded region. However, we do impose the restriction on $p(\Sigma)$ because the fixed points obtained without this restriction put a lot of probability mass on small values of Σ and compensate it by the large variance of B conditional on Σ. We find these priors not to be reasonable so an easy way to select reasonable behavior is to restrict the prior $p(\Sigma)$.

[29]To compute 'total output loss in the first 5 years' due to a monetary policy shock we sum the

importantly, the dynamics of output is mean-reverting, consistently with the long-run neutrality of money. Note, also, that the error bands are narrower in our posterior than with a flat prior, implying that we have incorporated useful information in the estimation.

The dashed line in panel D of Figure 4 plots a posterior corresponding to the fixed point with the highest entropy. It is comforting that this posterior confirms the main features of the highest marginal likelihood plotted with the solid line: higher persistence than OLS and mean reversion. As is well known, higher entropy is roughly related to higher dispersion, so it is intuitive that this fixed point shows larger posterior variance.

We report prior sensitivity analysis in the Online Appendix. We show that a range of reasonable priors on initial growth rates supports the conclusion that the response of output to a monetary policy shock is consistent with long-run neutrality of money. Moreover, most of these priors imply that the effect of a monetary shock is stronger and more persistent than in CEE, although the prior based on the data preceding the estimation sample is an exception here.

5 Conclusions

We have proposed using priors about observables and applied them to the estimation of Bayesian VARs. Priors about observables are easy to interpret and, as shown by our empirical applications, they often make a significant difference in empirical work.

To our knowledge we are the first to derive the posterior consistent with these priors in a formal way. We show the inverse problem that defines the prior on parameters that is consistent with a prior on observables, reformulate it as a fixed point problem,

median impulse response of the quarterly GDP in the first 5 years, and then divide by 4 in order to convert the result into annual GDP.

we give a numerical algorithm to find this fixed point and we show that this algorithm converges in the discrete case. This algorithm works even in very high-dimensional problems that we consider.

Application of Bayesian priors to VARs has obviously been a successful line of research. Standard priors on parameters such as those of Litterman, Sims and Zha have been useful in forecasting. But the specification of such priors is mostly experience-based and often not fully justified from a subjective point of view. Variants of these standard priors might give very different results and, as we show, might represent prior knowledge about observables that most economists would not hold. This presents serious problems when a researcher hopes that a VAR procedure will uncover unobservable features of the economy, such as e.g. impulse responses: if the stated prior does not represent the analysts' prior belief, the resulting posterior is not the best estimate of the unobservable quantities. In a way we advocate a 'more Bayesian' approach, providing a more natural representation of prior knowledge about the economy by focusing on observables.

Thus, the priors on observables we propose in this paper can serve as a cross-check on the standard priors and as an alternative to them.

Is it obvious how to formulate priors on observables? Certainly not. A researcher specifying a prior on observables needs to think hard about these observables and take a multitude of specification choices. In each of our two examples we used a different reasoning to arrive at the prior density and we do not doubt that many alternative reasonable priors could be constructed for these cases, possibly with different implications for the posterior. However, we contend that the approach we propose is more intuitive than the standard approach of specifying a prior about parameters directly. The possible different priors on observables can be evaluated much more intuitively as the issue is simply what is a best representation of our prior knowledge about

observables, for which most analysts do have a clear prior idea.

In any case the joint density of VAR parameters is a very high-dimensional object as well, and formulating it also requires lots of specification choices, 'weights' and 'shrinkage factors.' When thinking of the plausibility of these choices we are in the dark, because the VAR parameters are hard to interpret unless for their implications on observables. At the very least our work suggests that researchers using priors on parameters could use the accuracy test of section 3.3 to choose from among several alternative priors on parameters by examining their implications for observables.

Much future work remains. The empirical examples we have considered are mostly demonstrative and could be investigated further. Other ways of specifying priors on observables should be explored. Priors on observables could be used in many other applications and econometric models. The relation between 'reasonableness' of the prior on observables and the out-of-sample forecasting performance should be studied. Extending our analytical results would be useful. For example, our convergence result in Proposition 4 should be generalized in various directions, including the case of multiple solutions to the inverse problem and continuous distributions. Studying convergence when the fixed point problem does not have a solution may be useful in practice, as it may lead to systematic ways of modifying p_Y so as to guarantee existence.

Appendix A Proofs of Propositions 1 to 5 and Result 1

A first concern whenever we state results about $\mathcal{F}$ is to ensure that this mapping is well defined. This is not obvious since the expression contains a term dividing by $\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g$, a quantity that could be zero for some g's. To show that $\mathcal{F}$ is well

defined at g it is enough to check that $\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g > 0$ for all $\overline{Y}$ with $p_Y(\overline{Y}) > 0$.

Proof of Proposition 1

It is clear that for $g = p_\theta$ we have $\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g = p_Y(\overline{Y}) > 0$ so that $\mathcal{F}$ is well defined.

We have for all $\overline{\theta} \in \Theta$

$$\mathcal{F}(p_\theta)(\overline{\theta}) = \int_{\mathcal{Y}} p_{Y|\theta}(\overline{Y};\overline{\theta})\ p_\theta(\overline{\theta})\ d\overline{Y} = p_\theta(\overline{\theta}) \int_{\mathcal{Y}} p_{Y|\theta}(\cdot;\overline{\theta}) = p_\theta(\overline{\theta}).$$

The first equality holds from the definition of $\mathcal{F}$ and (4), the second equality takes $p_\theta(\overline{\theta})$ before the integral since it does not depend on $\overline{Y}$. The last equality holds because $p_{Y|\theta}(\cdot;\overline{\theta})$ is a probability density and therefore it integrates to 1 over $\mathcal{Y}$. ■

Proof of Proposition 2

Consider the continuous case. Let $p_\theta > 0$ be the solution of (4) considered in the statement of the proposition and consider $\widetilde{p}_\theta$ a possibly different solution of (4). Take $\delta(\theta) = \frac{\widetilde{p}_\theta(\theta)}{p_\theta(\theta)} - 1$, then, for all $\overline{Y}$ such that $p_Y(\overline{Y}) > 0$

$$E\left(\delta(\theta)|\, Y = \overline{Y}\right) = \int_\Theta \delta(\overline{\theta}) \frac{p_{Y|\theta}(\overline{Y};\overline{\theta}) p_\theta(\overline{\theta})}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ p_\theta}\ d\overline{\theta} = \int_\Theta \frac{p_{Y|\theta}(\overline{Y};\overline{\theta})\ \widetilde{p}_\theta(\overline{\theta})}{p_Y(\overline{Y})}\ d\overline{\theta} - 1 = 0. \tag{A.1}$$

The first equality uses $p_{\theta|Y}$ in terms of Bayes' formula and that we are considering $\overline{Y}$ such that $\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ p_\theta > 0$. The second and third equality use that $p_\theta, \widetilde{p}_\theta$ satisfy (4).

Since (A.1) holds a.s. in Y, completeness with respect to θ implies $\delta(\theta) = 0$, therefore $\widetilde{p}_\theta = p_\theta$ a.s. hence the solution is unique.■

Proof of Proposition 3

Consider the set $Y^0 \equiv \left\{\overline{Y} \in \mathcal{Y} : p_{Y|\theta}(\overline{Y};\cdot) = 0\right\}$. Equation (4) implies that if $\overline{Y} \in Y^0$ then $p_Y(\overline{Y}) = 0$ hence $Prob(Y \in Y^0) = \int_{Y^0} p_Y = 0$. Clearly, if $g > 0$ then $\int_\Theta p_{Y|\theta}(Y;\cdot)\ g > 0$ for all $\overline{Y} \notin Y^0$, so that $\mathcal{F}$ is well defined at g^* a.s. in Y.

At a fixed point we have $g^*(\theta) = \int_{\mathcal{Y}} \frac{p_{Y|\theta}(\overline{Y};\theta)g^*(\theta)}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g^*}\ p_Y(\overline{Y})\ d\overline{Y}$ for all θ. Given $g^* > 0$ we cancel $g^*(\theta)$ from both sides to have $1 = \int_{\mathcal{Y}} \frac{p_{Y|\theta}(\overline{Y};\theta)}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g^*}\ p_Y(\overline{Y})\ d\overline{Y}$ for all θ, and we have the second equality in

$$E\left(\frac{p_Y(Y)}{\int_\Theta p_{Y|\theta}(Y;\cdot)\ g^*}\middle|\ \theta\right) = \int_{\mathcal{Y}} \frac{p_{Y|\theta}(\overline{Y};\theta)}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g^*}\ p_Y(\overline{Y})\ d\overline{Y} = 1.$$

This holds a.s. in θ. Therefore, taking $\delta(Y) = \frac{p_Y(Y)}{\int_\Theta p_{Y|\theta}(Y;\cdot)\ g^*} - 1$, completeness implies $\delta(Y) = 0$ and that $\int_\Theta p_{Y|\theta}(Y;\cdot)\ g^* = p_Y(Y)$ almost surely in Y.■

Proof of Proposition 4

Extending our argument at the beginning of Proposition 3 to the discrete case, $\mathcal{F}(g)$ is well defined in a set $S \subset R^N_{++}$.

Taking derivatives of $\mathcal{F}$ mechanically we have for all $i, j = 1, ..., N$

$$\frac{\partial \mathcal{F}(g)_i}{\partial g_j} = \mathcal{I}_i(j) \int_{\mathcal{Y}} \frac{p_{Y|\theta}(\cdot;\theta_i)}{\sum_k p_{Y|\theta}(\cdot;\theta_k)\ g_k}\ p_Y - \int_{\mathcal{Y}} \frac{\left[p_{Y|\theta}(\cdot;\theta_i)\right]^2}{\left(\sum_k p_{Y|\theta}(\cdot;\theta_k)\ g_k\right)^2}\ p_Y\ g_i \qquad (A.2)$$

where $\mathcal{I}_i(j) = 1$ if $i = j$ and $\mathcal{I}_i(j) = 0$ $i \neq j$.

Let G^* be diagonal matrix with $p_{\theta,i}$ in the i-th diagonal entry, and $\left[p_{Y|\theta}(\overline{Y})\right]$ be the $N-$dimensional column vector with typical $i-$th element $p_{Y|\theta}(\overline{Y};\theta_i)$. Define

$$\Delta^* = \left[\int_{\mathcal{Y}} \left[p_{Y|\theta}(\overline{Y})\right]\ \ \left[p_{Y|\theta}(\overline{Y})\right]'\ p_Y(\overline{Y})^{-1}\ d\overline{Y}\ \right] G^*, \qquad (A.3)$$

with typical element $\Delta^*_{ij} = \int_{\mathcal{Y}} p_{Y|\theta}(\cdot;\theta_j)\ p_{Y|\theta}(\cdot;\theta_i)\ p_Y^{-1}\ p_{\theta,i}$.

Evaluating the right side of (A.2) at $g = p_\theta$, using (6) we find $\frac{\partial \mathcal{F}(p_\theta)_i}{\partial g_j} = \mathcal{I}_i(j) - \Delta^*_{ij}$ where $\mathcal{I}_i(j) = 1$ if $i = j$, $\mathcal{I}_i(j) = 0$ if $i \neq j$. Hence

$$\frac{\partial \mathcal{F}(p_\theta)}{\partial g'} = I - \Delta^*, \qquad (A.4)$$

where I is the identity matrix.

Denote the (possibly complex) eigenvalues of Δ^* by λ_n. We now show that for all $n = 1, ..., N$

$$\lambda_n \text{ is a real number and } 0 < \lambda_n \leq 1 \qquad (A.5)$$

It is easy to verify that the columns of Δ^* add up to 1. A well known result in matrix algebra guarantees that all eigenvalues satisfy $|\lambda_n| \leq 1$.

Given any vector $v \neq 0$, simple algebra gives

$$v'G^*\Delta^* v = \int_{\mathcal{Y}} \left(v'G^* \ \left[p_{Y|\theta}(\overline{Y})\right]\right)^2 p_Y(\overline{Y})^{-1} \ d\overline{Y}.$$

For a function δ defined as $\delta(\theta_i) = v_i$ we have $E(\delta(\theta) \mid Y = \overline{Y}) = v'G^* \left[p_{Y|\theta}(\overline{Y})\right] / p_Y(\overline{Y})$. Therefore

$$v'G^*\Delta^* v = \int_{\mathcal{Y}} \left[E(\delta(\theta) \mid Y)\right]^2 \ p_Y(\overline{Y}) \ d\overline{Y} > 0,$$

the inequality follows because completeness and $\delta(\theta_i) \neq 0$ imply $E(\delta(\theta) \mid Y) \neq 0$ with positive probability.

Therefore, $G^*\Delta^*$ is positive definite, hence all its eigenvalues are real and strictly positive. All that remains to be shown is that the eigenvalues of Δ^* inherit this property.

Obviously

$$\Delta^* = (G^*)^{-1} G^*\Delta^*.$$

Clearly $(G^*)^{-1}$ is symmetric and positive definite and we just proved that the same is true of $G^*\Delta^*$. The product of two symmetric and positive definite matrices has all eigenvalues real and strictly positive (e.g. this is a special case of Serre (2010) Proposition 6.1). Hence, we have shown that all eigenvalues λ_n are real and strictly positive. This ends the proof of (A.5).

Equation (A.4) implies that all eigenvalues of $\frac{\partial \mathcal{F}(g_\theta)}{\partial g'}$ are $1 - \lambda_n$, and, by (A.5) we have $|1 - \lambda_n| < 1 \ n = 1, ..., N$. Therefore, a standard argument implies that successive approximations on $\mathcal{F}$ locally converge to g_θ.■

Proof of Result 1

The result relies on the law of iterated expectations when (Y, θ) are distributed as $p^g p_Y$, that is, at the joint distribution that happens to emerge at this particular

iteration of the algorithm. To avoid confusion from having so many joint distributions on (Y, θ) we prove from scratch that

$$\begin{aligned} E_{\mathcal{F}(g)}(q(\theta)) &= \int_{\Theta} q(\overline{\theta}) \left(\int_{\mathcal{Y}} p^g(\overline{\theta}|\overline{Y}) \; p_Y(\overline{Y}) \; d\overline{Y} \right) d\overline{\theta} \\ &= \int_{\mathcal{Y}} \left(\int_{\Theta} q(\overline{\theta}) \; p^g(\overline{\theta}|\cdot) \; d\overline{\theta} \right) \; p_Y = E_{p_Y} \left(\; E_{p^g(\cdot|Y)}(q(\theta)) \; \right) \end{aligned}$$

The first equality above holds by definition of $\mathcal{F}(g)$, the second by Fubini's theorem and the third by definition of E_{p_Y}.■

Appendix B Standard priors for VARs

The flat (noninformative) prior is $p(B, \Sigma) \propto |\Sigma|^{-\frac{N+1}{2}}$, following e.g. Zellner (1971), Ch.8.

The remaining priors, 'Minnesota' prior, the 'Sims Zha (1998)' prior and the 'Dynare' prior, originate in Litterman (1979) and Doan et al. (1984). For reasons discussed in these and other papers, all these priors are centered at parameter values implying that the variables follow independent Random Walks, but they have different prior variances.

The functional form of the priors is Normal-Inverted Wishart form with parameters M, Q, S, v, see (11)-(12). All three priors use the same values of M, S, v and they differ only in the value of Q. The matrix M has 1s in the positions corresponding to the first own lag of each variable and 0s everywhere else, reflecting the postulate that the variables follow independent random walk models. We follow common rules of thumb when setting the remaining parameters. Namely, we set the parameters S, v using the 'empirical Bayes' approach. This approach is common practice and consists of the following steps. First, we estimate a univariate autoregression with P lags for each of the variables, using the estimation sample. Then we set S and v such that $E(\Sigma)$ is a diagonal matrix with the error variances of these univariate

autoregressions on the diagonal. We set the degree of freedom parameter to $v = 10$ in order to have a rather loose prior. Next, we build three versions of the parameter Q. The Q in the Minnesota prior approximates the prior of Litterman (1986) and follows the baseline recommendations of the RATS software manual (Doan, 2000). The Q in the Sims and Zha (1998) prior combines the Minnesota prior with the 'dummy observations prior' following Sims and Zha (1998). The Q in the Dynare prior also combines the Minnesota prior with the dummy observations prior but with somewhat different settings, namely with the settings used e.g. in Sims (2002) and implemented as the default in the Dynare software (Adjemian et al., 2011). In terms of Sims and Zha (1998) notation, in the the Minnesota prior we take $\lambda_1 = 0.2$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 10^5$, $\mu_5 = 0$, $\mu_6 = 0$; in the Sims and Zha (1998) prior we take $\lambda_1 = 0.2$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 1$, $\mu_5 = 1$, $\mu_6 = 1$; and in the Dynare prior we take $\lambda_1 = 0.33$, $\lambda_2 = 1$, $\lambda_3 = 0.5$, $\lambda_4 = 10^5$, $\mu_5 = 2$, $\mu_6 = 5$.

References

Adjemian, S., Bastani, H., Juillard, M., Mihoubi, F., Perendia, G., Ratto, M., and Villemot, S. (2011). Dynare: Reference manual, version 4. Dynare Working Papers 1, CEPREMAP.

Andrle, M. and Plasil, M. (2016). System priors for econometric time series. IMF Working Papers 16/231, International Monetary Fund.

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

Blanchard, O. and Perotti, R. (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *Quarterly Journal of Economics*, 117(4):1329–1368.

Bonhomme, S. and Robin, J.-M. (2010). Generalized non-parametric deconvolution with an application to earnings dynamics. *Review of Economic Studies*, 77(2):491–533.

Canay, I. A., Santos, A., and Shaikh, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559.

Carrasco, M. and Florens, J.-P. (2011). A spectral method for deconvolving a density. *Econometric Theory*, 27(03):546–581.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 6B of *Handbook of Econometrics*, chapter 77, pages 5633–5751. Elsevier.

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 1A*, pages 65–148. North-Holland, Amsterdam.

Christiano, L. J., Trabandt, M., and Walentin, K. (2011). Introducing financial frictions and unemployment into a small open economy model. *Journal of Economic Dynamics and Control*, 35(12):1999–2041.

Del Negro, M. and Schorfheide, F. (2004). Priors from general equilibrium models for VARs. *International Economic Review*, 45(2):643–673.

Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007). On the fit of New Keynesian models. *Journal of Business & Economic Statistics*, 25:123–143.

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.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, first edition.

Evans, G. W. and Honkapohja, S. (2002). *Learning and Expectations in Macroeconomics*. Princeton University Press, New York.

Florens, J.-P., Mouchart, M., and Rolin, J.-M. (1990). *Elements of Bayesian Statistics*. M. Dekker, New York.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2):436–451.

Giannone, D., Lenza, M., and Primiceri, G. E. (2018). Priors for the long run. *Journal of the American Statistical Association*. forthcoming.

Gupta, M. and Srivastava, S. (2010). Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy*, 12(4):818–843.

Ingram, B. F. and Whiteman, C. H. (1994). Supplanting the 'Minnesota' prior: Forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics*, 34(3):497–510.

Jarociński, M. and Lenza, M. (2018). An inflation-predicting measure of the output gap in the euro area. *Journal of Money, Credit and Banking*, 50(6):1189–1224.

Jarociński, M. and Marcet, A. (2010). Autoregressions in small samples, priors about observables and initial conditions. Working Paper 1263, European Central Bank.

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.

Litterman, R. B. (1979). Techniques of forecasting using vector autoregressions. Federal Reserve Bank of Minneapolis Working Paper number 115.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions - five years of experience. *Journal of Business and Economic Statistics*, 4(1):25–38.

Marcet, A. and Sargent, T. J. (1989). Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory*, 48(2):337–368.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhya : The Indian Journal of Statistics Series A*, 64(2):306–322.

Rubio-Ramírez, J. F., Waggoner, D. F., and Zha, T. (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *Review of Economic Studies*, 77(2):665–696.

Serre, D. (2010). *Matrices: Theory and Applications*. Springer, second edition.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Sims, C. A. (2002). The role of models and probabilities in the monetary policy process. *Brookings Papers on Economic Activity*, 33(2):1–62.

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

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.*, 34(4):1265–1269.

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.

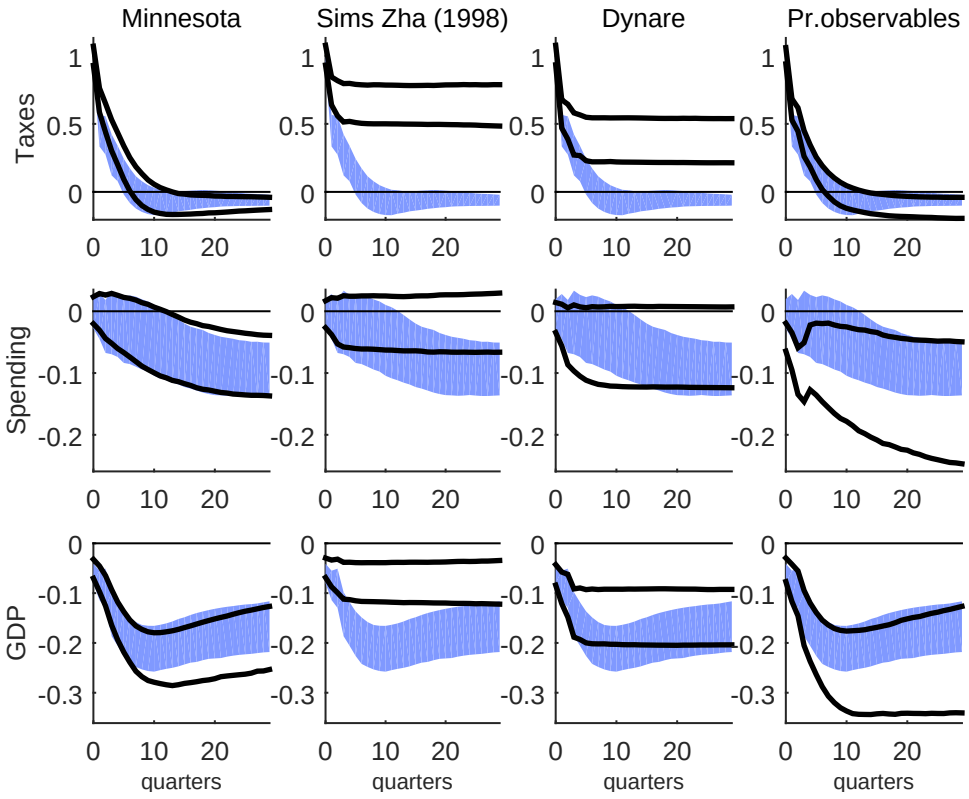

Figure 1 – Response to a Tax Shock: OLS estimation (shaded area, the same across columns) and Bayesian estimations using four informative priors. Y-axis gives quantiles 0.16 and 0.84 of the distribution of impulse response coefficients, in percent.

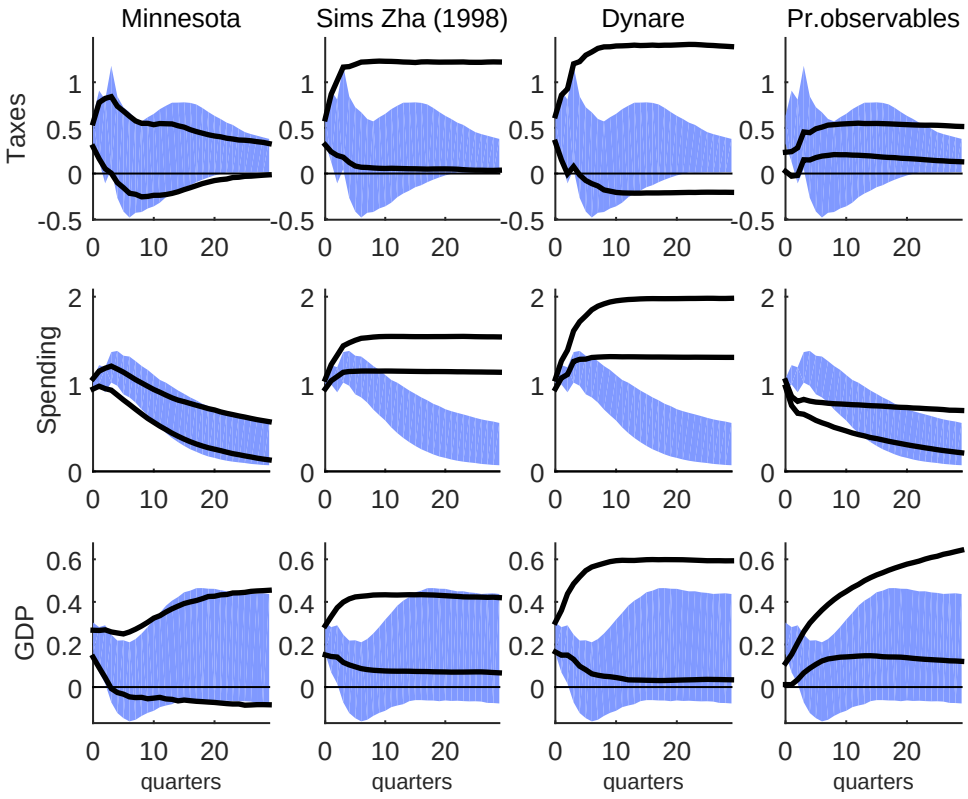

Figure 2 – Response to a Spending Shock: OLS estimation (shaded area, the same across columns) and Bayesian estimations using four informative priors. Y-axis gives quantiles 0.16 and 0.84 of the distribution of impulse response coefficients, in percent.

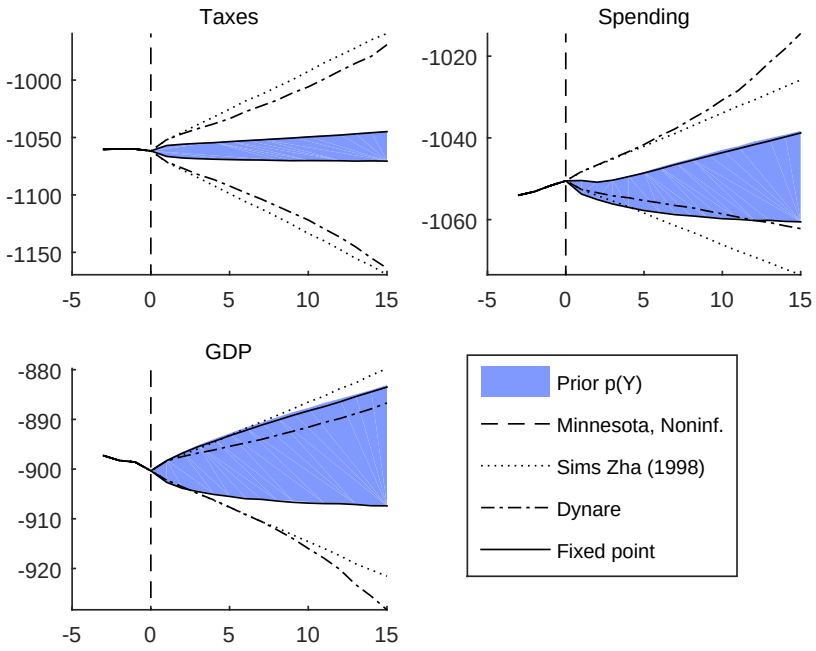

Figure 3 – Density of Taxes, Spending and GDP (in logs times 100) implied by alternative priors. Y-axis gives quantiles 0.05 and 0.95 of the distribution of each variable in periods 1 to 15 of the estimation sample.

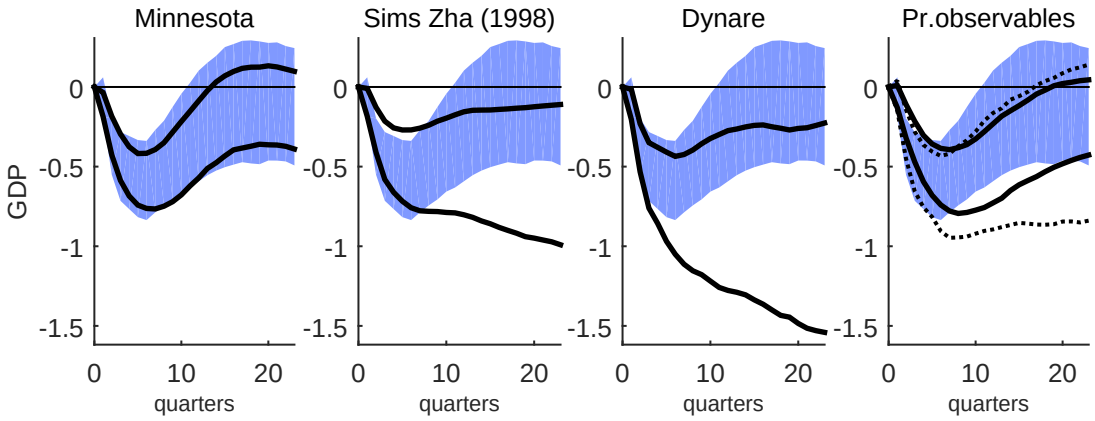

Figure 4 – Response of output to a monetary policy shock: OLS estimation (shaded area, the same across columns) and Bayesian estimations using four informative priors. Y-axis gives quantiles of 0.05 and 0.95 of the distribution of impulse response coefficients, in percent.

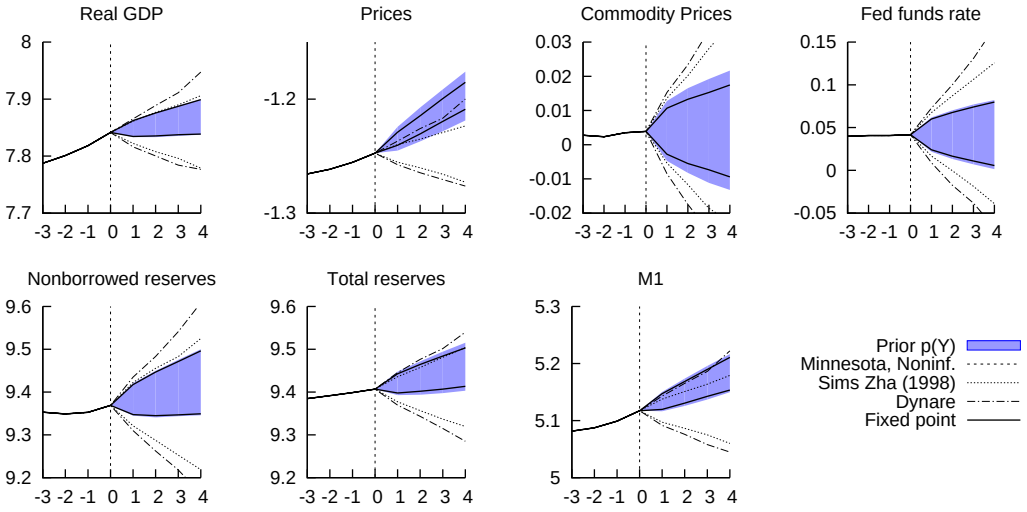

Figure 5 – Density of the observables implied by alternative priors. Y-axis gives quantiles 0.05 and 0.95 of the distribution of each variable in periods 1 to 4 of the estimation sample, in logs.