Priors about Observables in Vector Autoregressions*

Marek Jarociński
European Central Bank

Albert Marcet
Institut d'Anàlisi Econòmica CSIC,
ICREA, Barcelona GSE, MOVE, UAB

October 28, 2016

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. Our proposal is to use prior information on observables systematically. We show how this kind of prior can be used under strict probability theory principles. We state the inverse problem to be solved and we propose a numerical algorithm that works well in practical situations with a large number of parameters. We prove various convergence theorems for the algorithm. Using examples from the VAR literature, we show

*We thank Gianni Amisano, Manolo Arellano, Stéphane Bonhomme, Tony Braun, 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. All errors are our own. Albert Marcet acknowledges support from Axa Research Fund, DGES (Ministerio de Educación y Ciencia), CIRIT (Generalitat de Catalunya) 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.

how priors on observables can address a priori weaknesses of standard priors, serving as a cross check and an alternative formulation.

Keywords: Vector Autoregression, Bayesian Estimation, Prior about Observables, Inverse Problem, Monetary Policy Shocks

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] It is still the case that the literature offers a variety of priors on 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 often lack intuitive interpretation so it is difficult to claim that an analyst has proper priors about VAR parameters.[2]

[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 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 equation with the price level on the left hand side in a VAR, so an analyst is unlikely to have a subjective prior about it.

Moreover, we argue that priors on parameters that are standard in the literature can not represent the beliefs about observables that analysts do have: when we derive the prior on observables that is implied by various standard priors on parameters we find very disparate behavior of observables and often prior beliefs that a reasonable analyst would never have. In this case the resulting posterior lacks a proper Bayesian justification.

Given that economists do have priors about the behavior of observable variables our proposal is to be 'more Bayesian' in the estimation of VARs and to incorporate this prior knowledge in the estimation.

At the very least, Bayesian VAR applications should examine if the used prior on parameters implies a reasonable prior behavior for observables. Our main proposal, however, is to incorporate prior information in a direct way, namely, to state explicitly a prior on observable variables and to obtain the posterior consistent with it. This can be done as follows. First, 'translate' the prior on observables to an equivalent prior on parameters. This is found 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 use successive approximations on this mapping to compute the fixed point. We prove that under mild assumptions the fixed point condition is necessary and sufficient for the solution and that successive approximations converge locally to the solution. Finally, we propose an approximate conjugate algorithm that speeds up the computation of the fixed point and of the posterior.

Using priors on observables is not only a natural way to incorporate information that analysts do have, it also focuses the discussion about what is a 'good' prior on the observables and, therefore, differences in priors are easier to interpret. Different prior opinions about, say, the variance of output growth rate, are not going to be

huge across analysts and they are easy to interpret.

Our algorithm and fixed point formulation is applicable to any statistical model, but in this paper we focus on applications to the structural Bayesian VARs.

To show that our approach works in practical applications we use it to reexamine three important VAR studies: the estimation of fiscal policy effects in Blanchard and Perotti (2002), the estimation of monetary policy effects in Christiano et al. (1999) and in Romer and Romer (2004). 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, Sims and Zha.[3] We show that these standard VAR priors on parameters actually imply widely disparate priors on observables. Some of them imply 'crazy' behavior of observables, a prior knowledge that no reasonable economist or policy maker would hold, therefore these standard priors are not justified from a subjective Bayesian point of view.

Our approach incorporates into the prior the knowledge about the economy that economists may have. We find that incorporating such knowledge does matter for the results. With subjective priors on observables we find larger fiscal multipliers than Blanchard and Perotti (2002), more persistent real effects of monetary policy than Christiano et al. (1999), and highlight a mismatch between the Romer and Romer (2004) evidence on the effects of monetary policy and a standard New Keynesian model.

These examples show, first, that a prior on observables may be useful in clarifying empirical results, as they eliminate some of the inconsistencies that priors on parameters generate. Second, it reduces posterior variance relative to the noninformative prior by incorporating useful information in the inference. Third, our algorithm works well in practice even in a relatively large VAR where the fixed point we compute has

[3]See Litterman (1986); Sims and Zha (1998); Sims (2002).

hundreds of parameters.

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 ways of doing so, as we construct priors on observables in a different way for each example: we use knowledge about the economy stated by Blanchard and Perotti in the first example, we use an empirical Bayes prior in the second example, and a prior based on a structural model of the economy (related to the priors in Del Negro and Schorfheide 2004, and others) in the third example.

Another advantage of our prior is that it produces good results when evaluated from a classical perspective. In Jarociński and Marcet (2010) we show that it reduces the mean squared error relative to the various classical small sample bias correction techniques considered.

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

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., 2016), DSGE model-based priors (Ingram and Whiteman 1994, Del Negro and Schorfheide 2004, Del Negro et al. 2007, Christiano et al. 2011 and others). 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]

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]

[4]For example, the DSGE-model-based priors mentioned above in effect do not solve the inverse problem described in section 2. In light of our results, they can be justified as performing only 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

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-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

avoids any matrix inversion.

[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.

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 and 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 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.[9]

We stress that our fixed-point approach to solving the inverse problem is not specific to VARs, it may be used for handling priors on observables in other models.

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 directly on parameters p_θ. 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 this prior information we first translate the prior on observables p_Y 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

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

[9] In the current paper we discuss some of these results in section 3.2 and footnote 12.

obtain the posterior that is consistent with the prior on observables.

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. One may 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 is just stating the prior idea that the researcher holds about the behavior of Δy_1, notice that it 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.

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

$$\mu_\Delta = E(\Delta y_1) = E(\alpha + (\rho - 1)y_0)$$
$$\sigma_\Delta^2 = \text{Var}(\Delta y_1) = \text{Var}(\alpha + (\rho - 1)y_0) + \sigma_\epsilon^2$$

and 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 $p_{Y|\theta}$ 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.'

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 like (4) can rarely be solved analytically.[10] We now reformulate our inverse problem in terms of a fixed point problem that facilitates computation. We first present some results on necessity and sufficiency of the fixed point condition that hold in the case of continuous distributions such as those used in VAR applications. We propose an algorithm for finding the fixed point by successive approximations. We prove that this algorithm converges for the discrete case and we show that the continuous case can be approximated by discretizing appropriately. Finally, we describe the approximate conjugate fixed point iteration that we use in practice and we show how to check for accuracy.

[10] 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 just generalizing to a prior on the growth rates in two periods, $t = 1, 2$, yields a problem where parameters enter non-linearly and an analytic solution is not available. The change of variable formula does not help either, see the online Appendix G for a further discussion.

3.1 The continuous case

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 $\mathcal{F}$:

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

Let us comment on the notation. First, we have written the integrals in (4) and in (5) in terms of densities, we consider the discrete case in section 3.2. Second, 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: the term $\frac{p_{Y|\theta}(\overline{Y};\overline{\theta})\ g(\overline{\theta})}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g}$ is the posterior distribution obtained when the prior on parameters is given by g and when the data realization $\overline{Y}$ is observed. Therefore, $\mathcal{F}(g)$ is a mixture of posteriors for different realizations $\overline{Y}$, each weighted by its density $p_Y(\overline{Y})$.

Clearly, we have $\mathcal{F}(g) \geq 0$ and if $\mathcal{F}(g)$ is integrable then Fubini implies $\int_\Theta \mathcal{F}(g) = 1$, so that $\mathcal{F}(g)$ is itself a density.

We now show that there is a close relation between solutions of (4) and fixed points of the mapping $\mathcal{F}$. First of all we prove

Proposition 1. ***(Necessity)*** *If p_θ satisfies (4) and $p_Y(\overline{Y}) > 0$, then $\mathcal{F}(p_\theta)$ is well defined and p_θ is a fixed point of $\mathcal{F}$.*

Even though necessity is obtained under very general conditions it turns out that uniqueness of solutions to (4) and sufficiency of a fixed point condition $\mathcal{F}(g^*) = g^*$ are closely related through the concept of *completeness*, defined as follows.

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 all $b \in \mathcal{B}$ then $\delta = 0$ a.s. in A.*

The completeness conditions we will need essentially mean that the model $p_{Y|\theta}$ is identified, in other words that values of Y a.s. carry all relevant information about the value of θ and vice versa. The relationship between completeness and identification has been the object of recent research in non-parametric estimation, starting with Newey and Powell (2003). The discrete case analysed in the next section clarifies that all that is needed in the discrete case is that a matrix version of the likelihood $p_{Y|\theta}$ is invertible.

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

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

It follows from Propositions 1, 2 and 3 that if $p_{\theta,Y}$ is complete both with respect to Y and θ, and there exists a positive solution to (4), then the set of solutions to (4) and the set of positive fixed points of $\mathcal{F}$ coincide and this set is a singleton.

The above propositions suggest that instead of trying to solve problem (4) directly we can search for fixed points of the mapping $\mathcal{F}$.

Let us state, for future reference, 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) Start with g^0, an initial density of θ. 2) Given g^{z-1} find $g^z = \mathcal{F}(g^{z-1})$ for $z = 1, 2, ...$ Repeat 2) until convergence.*

Algorithm 1 avoids many difficulties often found when solving inverse problems. First, inversion of large matrices is entirely avoided. Second, g^z is guaranteed to be a

proper density at every iteration z, and thus one does not have to restrict the solution to be positive and to add up to 1.

It will turn out, however, that the restriction to positive fixed point $g^* > 0$ is important. It is possible to see that $\mathcal{F}$ has 'false fixed points' such that $g^{**} = \mathcal{F}(g^{**})$ where $g^{**} = 0$ for some θ and 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. The existence of such 'false fixed points' is trivial in the discrete case so we show some specific examples to the next section.

In section 3.4 we propose a practical approximation to Algorithm 1 that is likely to act as a regularization and that works well in the large VARs that are used in the literature. Furthermore, this approximate algorithm gives a positive density for θ everywhere, therefore it stays away from false fixed points.

Even though VAR applications usually consider continuous distributions of AR parameters, continuous densities have to be projected into classes of functions of finite elements. A discrete approximation is a good candidate for such a family. Indeed, in section 3.3 we discuss how to approximate continuous densities with discrete distributions. Thus we now move to the discrete case.

3.2 The discrete case

We now discuss discrete distributions of θ and Y. The discrete case is useful because it allows us to analytically prove convergence of the successive approximations algorithm using results in matrix algebra.

Assume that Y and θ are discrete variables that each take N possible values, that is $\mathcal{Y} = \left\{\overline{Y}_1, ..., \overline{Y}_N\right\}$ and $\Theta = \left\{\overline{\theta}_1, ..., \overline{\theta}_N\right\}$ for a finite integer N. The likelihood function is known and given by an $N \times N$ matrix Π with a typical element $\pi_{ij} = p_{Y|\theta}(\overline{Y}_j; \overline{\theta}_i)$. In this section p_Y is an N-dimensional vector that contains $p_Y(\overline{Y}_j)$ in the j-th element.

We write $g(\overline{\theta}_i) = g_i$. In the discrete case equation (4) specializes to

$$\Pi' g_\theta = p_Y \tag{6}$$

for some discrete distribution g_θ. Note that we use g for general probability vectors of θ and g_θ denotes the solution to (6).

We assume throughout that Π is invertible. Some results could be obtained without this assumption but not much more generality would be gained.

A quick look at (6) may suggest that solving inverse problems in discrete problems is an easy task, as it can be achieved by simply inverting the matrix Π'. However, in practice Π' is often large dimensional and ill-conditioned, making matrix inversion unfeasible. It is well known that this approach leads to ill-defined solutions where for an approximate Π' the resulting solution $g_\theta = (\Pi')^{-1} p_Y$ has some negative elements. In contrast, the algorithm of successive approximations on $\mathcal{F}$ completely sidesteps any matrix inversion. This plus the use of a conjugate approximate algorithm in section 3.4 below enables us to solve very high-dimensional problems.

The discrete version of the mapping $\mathcal{F}$ (5) is

$$\mathcal{F}(g)_i \equiv \sum_j \frac{\pi_{ij} g_i}{\sum_k \pi_{kj} g_k} \, p_Y(\overline{Y}_j) \quad \text{for all } i = 1, ..., N. \tag{7}$$

The issue of *existence* of a distribution g_θ that solves (6) is straightforward. Since Π is invertible $g_\theta = (\Pi')^{-1} p_Y$ is a well defined vector and it satisfies $\sum_{i=1}^N g_{\theta,i} = 1$.[11] But for existence we still need to assume that $(\Pi')^{-1} p_Y$ has only non-negative elements.

A trivial adaptation of Proposition 1 guarantees that if $p_Y > 0$, a probability vector g_θ that solves (6) is a fixed point of $\mathcal{F}$ (necessity). The following proposition guarantees sufficiency.

[11] This is because since Π has an eigenvector equal to $\mathbf{1}$ (a column vector with all elements equal to 1), we have $\mathbf{1}' g_\theta = \mathbf{1}' (\Pi')^{-1} p_Y = 1$.

Proposition 4. ***(Sufficiency, discrete case)*** *Assume that i)* Π *is invertible and ii)* g^* *is a fixed point of* $\mathcal{F}$ *such that* $g_i^* > 0$ *for all* $i = 1, ..., N$*. Then* g^* *is the unique solution of (6).*

The requirement that a fixed point satisfies $g^* > 0$ is not a technicality, it is important for sufficiency: there are indeed fixed points of $\mathcal{F}$ with some elements of g equal to zero which are not solutions to the inverse problem. In particular, it is easy to check that there is always a 'false fixed point' with $g_{\bar{i}}^{**} = 1$ for any $\bar{i}$. Also, fixing $g_{\bar{i}}^{**} = 0$ for some $\bar{i}$ gives $N - 1$ remaining equations and unknowns to find values for the remaining coordinates g_i^* $i \neq \bar{i}$ that satisfy the fixed point condition.

The following proposition guarantees that the successive approximation algorithm is locally stable under some conditions.

Proposition 5. ***(Convergence)*** *Assume that i)* Π *is invertible, ii) the vector* $g_\theta = (\Pi')^{-1} p_Y$ *satisfies* $g_{\theta,i} > 0$ *for all* i*. Then all eigenvalues of the derivative* $\frac{\partial \mathcal{F}(g_\theta)}{\partial g'}$ *are real and they belong to the interval* $[0, 1)$*.*

Therefore, successive approximations on $\mathcal{F}$ *converge locally to* g_θ*. Formally, 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* g_θ *such that for all* $g^0 \in S$ *we have* $g^z \to g_\theta$ *as* $z \to \infty$*.*

Let us discuss the above assumptions. Invertibility of Π is related to completeness and identification of the model $p_{Y|\theta}$. For example, if invertibility failed because two rows of Π were equal, this would mean that two different values of θ imply the same behavior of Y so that the likelihood $p_{Y|\theta}$ would not allow identification of θ.

Assumption ii) only adds a strict inequality $g_{\theta,i} > 0$ over and above the non-negativity requirement that is already needed for existence of a solution to (6). Given existence this strict inequality is a mild requirement. It is clear that the set of Π's and p_Y's for which $g_\theta = (\Pi')^{-1} p_Y$ is non-negative and it violates *ii)* is of measure zero.

This justifies using Algorithm 1 to solve the inverse problem. Under the assumptions, if the algorithm converges to a strictly positive g we can be certain that this is the solution to the inverse problem. Furthermore, if the solution of the inverse problem is strictly positive and we start the iterations sufficiently close, the algorithm will converge. Using homotopy, this ensures that we can always approximate the solution of the inverse problem with Algorithm 1.

As mentioned before, the iterations have to be kept away from 'false' fixed points. Since our algorithm relies on local convergence we can always use homotopy to build good initial conditions in a systematic way so as to stay within a neighborhood of the correct fixed point.[12] The conjugate approximate algorithm that we use in the empirical applications ensures that g^* is everywhere positive by construction.

3.3 A discrete approximation to the continuous case

When continuous densities of θ and Y exist the solution p_θ has to be approximated numerically by a class of functions with finite elements. In this subsection we show how this can be done using step functions, thus mapping the continuous case into the

[12]Some results in the literature state global convergence for the algorithm of Newton (2002), for example Martin and Ghosh (2008). But in fact these results do not accurately reflect the behavior of that algorithm. First, it is obvious that Newton's algorithm is not globally stable in the space of distributions because if the initial condition is set equal to one of the 'false' fixed points described in the text the algorithm stays there forever. Newton's algorithm should be re-designed to exclude these false fixed points and convergence proofs should be adapted. Second, it can be shown that in the vicinity of such points Newton's algorithm moves particularly slowly, therefore these 'false' fixed points slow down convergence. Third, combining results from stochastic approximation and our Proposition 5 one can show that Newton's algorithm converges asymptotically at a rate slower than $\sqrt{T}$ for most applications. On the other hand, applying our approach to non-parametric estimation alleviates or completely corrects these problems and, in particular, $\sqrt{T}$ convergence obtains. A formal proof of the statements in this footnote is available from the authors.

discrete distribution described in section 3.2. We find conditions guaranteeing that the fixed points of this modified problem converge to a solution of the continuous inverse equation (4) as the step-size becomes finer. Combining this result with Proposition 5 we can state that for sufficiently many iterations on $\mathcal{F}$ and sufficiently small step size ε we can approximate the continuous p_θ that solves (4) arbitrarily well.

Appendix B describes in detail how to partition $\mathcal{Y}$ and Θ each into $N_\varepsilon < \infty$ non-overlapping intervals denoted $\mathbf{Y}_j^\varepsilon$ and $\boldsymbol{\theta}_i^\varepsilon$ $i, j = 1, ..., N_\varepsilon$ respectively with interval width $\varepsilon > 0$. We discretize p_Y by an N_ε-dimensional probability vector p_Y^ε with elements $p_{Y,j}^\varepsilon = \int_{\mathbf{Y}_j^\varepsilon} p_Y$. We discretize $p_{Y|\theta}$ by an $N_\varepsilon \times N_\varepsilon$ matrix Π^ε with typical element $\pi_{i,j}^\varepsilon = \int_{\mathbf{Y}_j^\varepsilon \times \boldsymbol{\theta}_i^\varepsilon} p_{Y|\theta}$.

Let $g_\theta^\varepsilon \in R^{N_\varepsilon}$ be a discrete distribution that satisfies the discrete inverse equation for this approximation

$$\Pi^{\varepsilon\prime} g_\theta^\varepsilon = p_Y^\varepsilon \tag{8}$$

and let G_θ^ε be the corresponding cumulative distribution function $\int_{\boldsymbol{\theta}_j^\varepsilon} dG_\theta^\varepsilon = g_{\theta,j}^\varepsilon$ for all $j = 1, ..., N_\varepsilon$.

Proposition 6. ***(Approximation by step functions)*** *If the (continuous) inverse equation (4) has a unique solution density p_θ with a corresponding cdf G_θ, and the assumptions of Lemma 1 hold, then $G_\theta^\varepsilon \to G_\theta$ weakly as $\varepsilon \to 0$.*

The proof is in Appendix B.

3.4 Approximate conjugate algorithm

Proposition 5 is useful because it shows a precise sense in which successive approximations converge in the discrete case. But after experimenting with such discretizations

we found them impractical. The reason is that discretizing a likelihood function with very many parameters becomes highly costly computationally.[13]

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. In this algorithm, at each iteration we restrict the density g to be in a given parametric family that is conjugate with the likelihood. The 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. Therefore, after stating the algorithm we discuss how to check ex-post if the accuracy of the approximation is acceptable.

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 $g \in \mathcal{G}$.[14]

Algorithm 2. ***(Approximate conjugate algorithm)***

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 $E_{\mathcal{F}(g^z)}(q(\theta))$.

[13]For example, the VAR of Christiano et al. (1999) that we discuss in section 4.2 has 231 parameters. If we use, say, 10 intervals per parameter we would have $N_\varepsilon = 10^{231}$. This is much larger than the distance between the earth and the nearest star outside the solar system measured in millimeters (roughly, a mere $4{\cdot}10^{19}$mm.)

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

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. Let $p^g(\overline{\theta}|\overline{Y}) = \frac{p_{Y|\theta}(\overline{Y};\overline{\theta})\ g(\overline{\theta})}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g}$ denote the posterior distribution of θ obtained with the prior distribution g and given data realization $\overline{Y}$.

Result 1. [15] *Given any density g, for any function $q : \Theta \to R^\nu$ we have*

$$E_{\mathcal{F}(g)}(q(\theta)) = E_{p_Y}\left[E_{p^g(\cdot|Y)}(q(\theta))\right]. \tag{9}$$

This result suggests that the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$ required in Step 2 above can be computed using the following Monte Carlo procedure: draw J realizations of Y from p_Y, then split Step 2 into two steps: 2a) For each draw $\overline{Y}$ 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 (9) by averaging the posterior moments obtained in Step 2a over the J draws. The key is that if $\mathcal{G}$ is a family of conjugate priors for $p_{Y|\theta}$ then the moments computed in Step 2a are available in closed form so that this computation can be done very efficiently. When $\mathcal{G}$ is not conjugate then Algorithm 2 also works, but it 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 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. Consider a prior on observables p_Y describing the behavior of (y_1, y_2), hence an analytic solution is not available (see footnote 10). 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

[15] This result follows from the law of iterated expectations at the fixed point, but for arbitrary g $\mathcal{F}_{p_Y}(g)$ is not the marginal density of θ consistent with p_Y and $p^g_{\theta|Y}$, and thus we offer a (rather simple) proof of (9) in the Appendix.

of θ. Given a sample $\overline{Y} = (\overline{y}_1, \overline{y}_2)$ a standard result in Bayesian statistics fully characterizes the posterior as

$$(M^{po}, \mathcal{V}^{po}) = F_{\mathcal{N}}(M^{pri}, \mathcal{V}^{pri}; \overline{Y}) \tag{10}$$

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

Algorithm 3. ***(Approximate conjugate algorithm under normality)***

Let $\mathcal{G}$ 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 $\mathcal{V}^z$.[16]

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

Algorithm 3 shows how Algorithm 2 and Result 1 can be combined in a simple case.[17] But Algorithm 3 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

[16]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 3 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.

[17]For an application see Jarociński and Lenza (2016) pp. 24-26.

uncertainty about the innovation variance and generalize to the case of a multivariate model, a 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. This is what we do in our applications in section 4. Here is a full description of this algorithm.

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

$$y_t = \sum_{p=1}^{P} B_p \, y_{t-p} + c + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma), \quad t = 1, ..., T. \tag{11}$$

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

$$p(\text{vec}\, B|\Sigma) = \mathcal{N}(\text{vec}\, M, \Sigma \otimes Q), \tag{12}$$

$$p(\Sigma) = \mathcal{IW}(S, v), \tag{13}$$

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

As in Algorithm 3 we denote $M = E(B)$ and $\mathcal{V} = E(\text{vec}\, B(\text{vec}\, B)')$. We also denote the moments of Σ^{-1} as $D = E(\Sigma^{-1})$ and $\mathcal{H} = \text{diag}\, E(\text{vec}\, \Sigma^{-1} \, (\text{vec}\, \Sigma^{-1})')$. Analogous to (10), 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

$$(M^{po}, \mathcal{V}^{po}, D^{po}, \mathcal{H}^{po}) = F_{\mathcal{NIW}}(M^{pri}, Q^{pri}, S^{pri}, v^{pri}; \overline{Y}) \tag{14}$$

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 4. ***(Approximate conjugate algorithm for a Normal-Inverted Wishart prior in a VAR)***

Let $\mathcal{G}$ 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 3 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, under a certain choice of the objective function. We derive closed form expressions for M^z, Q^z, S^z, v^z in the Online Appendix. There are many ways one can approximate $\mathcal{F}(g^{z-1})$ with a Normal-Inverted Wishart density, but our experience suggests that the precise modelling choices there are not critical for the properties of the algorithm, so we go for the simplicity of implementation.

3.5 Accuracy

After performing the iterations we need to check the accuracy of the approximate solution 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 prior algorithm as described in the previous subsection. But g^Z does satisfy (4) exactly for a certain distribution of Y: namely, we can always plug g^Z in the left side of (4) and obtain the corresponding distribution for observables $p_Y^Z = \int_\Theta p_{Y|\theta}\, g^Z$. Since the prior densities p_Y that a researcher may state for observables can only be indicative, if p_Y^Z is 'reasonable close' to p_Y then g^Z should be an acceptable translation of p_Y.

With this motivation we check accuracy by comparing p_Y^Z and p_Y. 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 (blue shaded area) and the quantiles of the distribution of the observables implied by the approximate fixed point (solid line).

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 three applications of priors on observables to the estimation of structural VARs. All three 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) and Example 3 is a study on the same topic by Romer and Romer (2004).

The aim of the section is to show, first, that different standard priors on parameters available in the literature give significantly different results, and that there are few reasons a-priori to choose from among these alternatives. Second, we show that some of these priors on parameters imply a prior on observables that can not possibly represent prior knowledge of the analyst, therefore it is unjustified to use the resulting posterior from a Bayesian point of view. Third, we find that the algorithm proposed in section 3.4 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, in the second example we use an empirical Bayes prior, and in the third we use a structural economic model. We show that the prior on observables does affect the results and gives useful guidance, in fact changing considerably the empirical outcome relative to some published results. To the extent that this prior on observables may be a better representations of the analyst's prior knowledge we contend that the resulting posterior is better justified from a Bayesian point of view.

We first show the results from four standard priors for θ used commonly in the VAR literature. The first one is the flat (noninformative) prior, where the posterior mean of B is the OLS estimate. All three papers from which we take our examples, Blanchard and Perotti (2002) Christiano et al. (1999) and Romer and Romer (2004), use OLS estimation, hence the flat prior comes closest to replicating their results (apart from small discrepancies between their bootstrap and our Bayesian uncertainty bands). We add to that three informative priors for VARs in the Litterman, Sims and Zha tradition, using three off-the-shelf choices for the settings of these priors.

We refer to them respectively as 'Minnesota' prior (the default prior in the RATS computer package), 'Sims Zha (1998)' prior (a widely used version of the prior) and 'Dynare' prior (the default prior in the Dynare computer package). See Appendix C for precise definitions of these priors. We find that in two of the three applications these standard priors produce very disparate results. Moreover, many of them could be easily ruled out because they imply priors about observables that are completely unreasonable. Although sometimes one of these standard priors happens to produce results close to the results we obtain with our prior on observables, it is hard to predict ex ante for which prior on parameters this will occur in a given application.

We then apply our approach to each example. In each application we choose a different way to specify the prior on observables, appealing to different approaches used in Bayesian econometrics: i) the prior for Example 1 is formulated based on the experts' view about observables, as expressed by Blanchard and Perotti in their paper, ii) we apply an empirical Bayes prior to Example 2,[18] iii) we apply a prior based on a structural economic model to Example 3. Notice that in all three cases 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 3.4. Finally, we use Bayes' theorem in the standard way to compute the posterior.

In all three examples the VARs are specified in levels and the variables entering them are clearly nonstationary. Therefore, the prior density of these variables must be conditional on some initial state. A natural choice is to use as the initial state

[18]Empirical priors are contentious but they have been used in the literature nevertheless, and they have the virtue of transparency.

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 all the three examples: we present the empirical application, show results obtained with 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 and the estimation sample is 1960Q1-1997Q4.[19] 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 the quantiles 0.16 and 0.84 of the posterior distributions of the impulse responses.

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

The variables are 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.[20] 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

[20]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.

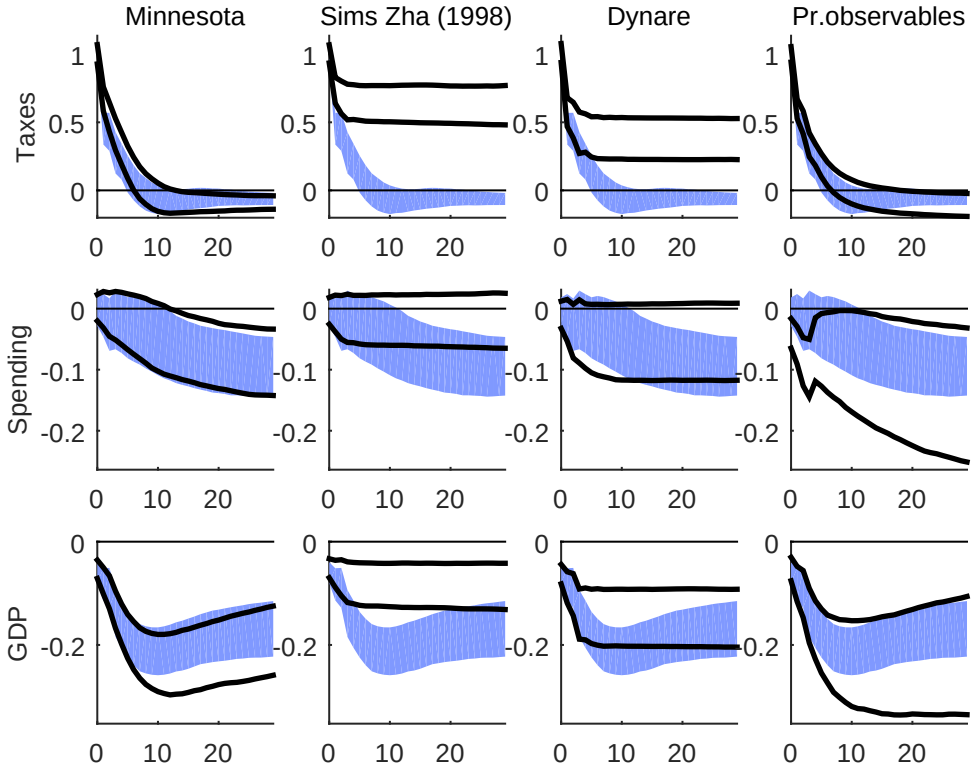

Figure 1 – Response to a Tax Shock: OLS estimation (shaded area, the same across columns) and Bayesian estimations using four informative priors. Quantiles 0.16 and 0.85.

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

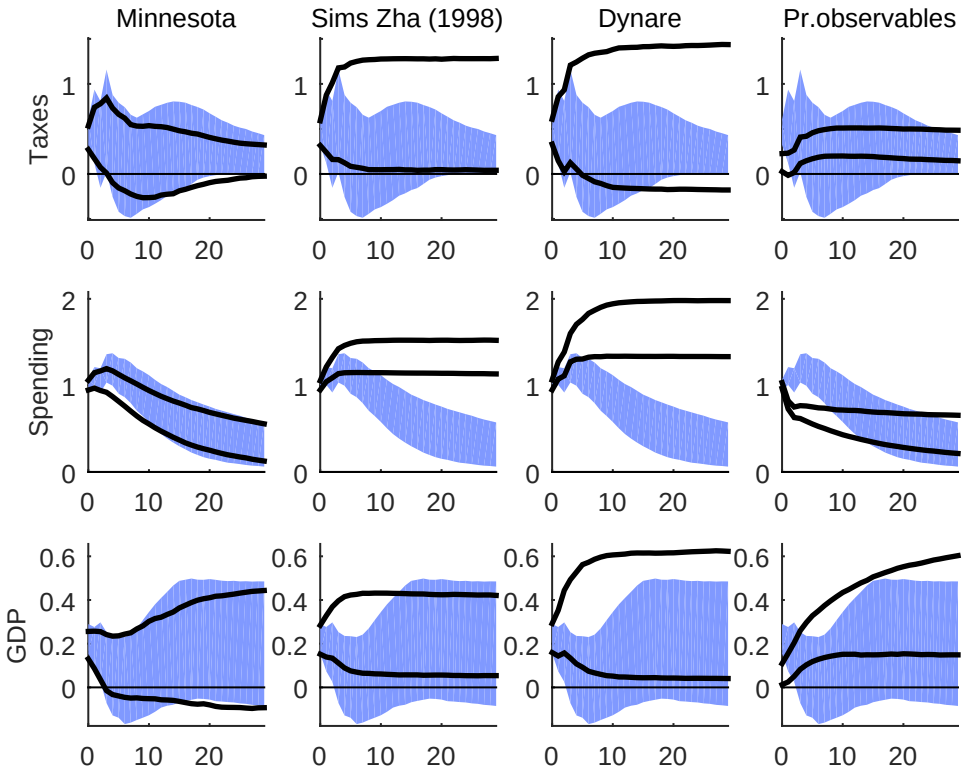

Figure 2 – Response to a Spending Shock: OLS estimation (shaded area, the same across columns) and Bayesian estimations using four informative priors. Quantiles 0.16 and 0.85.

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).[21] 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.

This figure is meant to show that standard priors on parameters cannot represent the opinion of the analyst in this application. Hence, the posteriors found with these prior distributions are not valid on 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

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

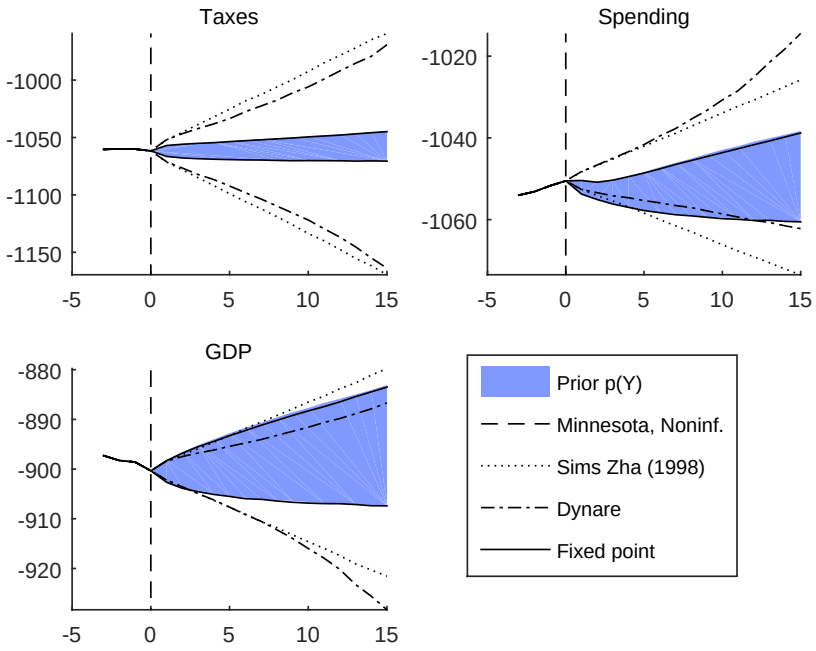

Figure 3 – Density of Taxes, Spending and GDP implied by alternative priors. Quantiles 0.05 and 0.95 of the distribution in periods 1-15 of the estimation sample.

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 Blanchard and Perotti (2002). One aspect of these priors is cointegration, for an alternative approach to priors about cointegration in VARs see Giannone et al. (2016).

The data that inform our prior are on real GDP for the period 1947-1960 and on taxes and spending in 1960.[22] We fit an AR(2) model into the GDP data for

[22]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.

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

$$\tau_t = \delta_x + \tau_{t-1} - \beta^\tau(\tau_{t-1} - x_{t-1} - c^\tau) + a_1 u_t^x + \sigma^\tau \varepsilon_t^\tau, \tag{15a}$$

$$g_t = \delta_x + g_{t-1} - \beta^g(g_{t-1} - x_{t-1} - c^g) + \sigma^g \varepsilon_t^g, \tag{15b}$$

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 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, as then 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. $N(NP+1) + N(N+1)/2 = 45$). 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 3.4 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. Figure 4 presents the evolution of two of the parameters of the Normal-Inverted Wishart density as we iterate from ten different starting points. We plot the top-left elements of matrices M and S, denoted $M(1,1)$ and $S(1,1)$ respectively. The third plot shows the evolution of the Kullback-Leibler divergence between $p(Y)$ and $\int_\Theta p(Y|\theta)\ g^z(\theta)d\theta$ (the left-hand side and the right-hand side of equation (4)) estimated from a sample of 1000 draws from each density.[23] We can see that after about 200 iterations the K-L divergence reaches the vicinity of zero (thereafter it fluctuates because of the estimation noise). 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.

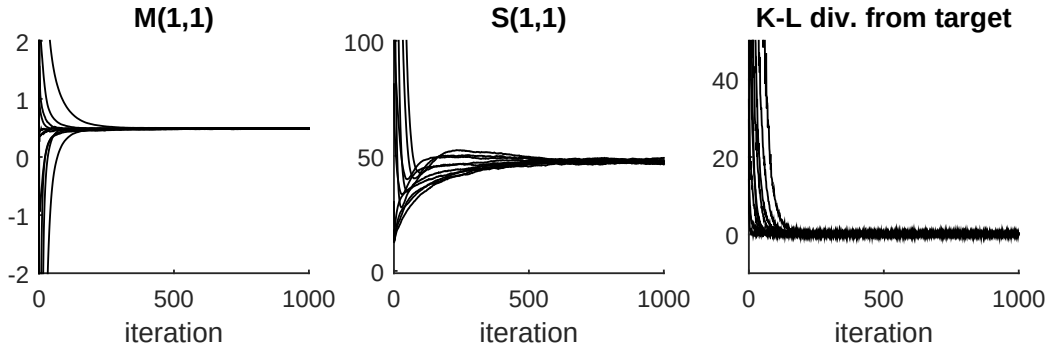

Figure 4 – Two parameters of g^z along the iterations and the estimated Kullback-Leibler divergence between $p(Y)$ and $\int_\Theta p(Y|\theta)\ g^z(\theta)d\theta$ along the iterations.

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

[23] We use $p(Y)$ as the weighting function in Kullback-Leibler divergence, i.e., we estimate $\int_\mathcal{Y} p(Y) \log\left(p(Y)\ /\ \int_\Theta p(Y|\theta)\ g^z(\theta)d\theta\right) dY$. We compute the estimate using the TIM package for Matlab (Rutanen, 2011), which implements the nearest-neighbor estimator the Kullback-Leibler divergence proposed by Wang et al. (2009).

with the shaded region the quantiles according to the prior on observables. The solid 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 0.4% after 12 quarters, compared with about 0.2% according to the OLS estimation, Minnesota, and Sims and Zha (1998) priors. The response of output obtained with the Dynare prior is close to 0.4% after 12 quarters, but it is associated with a much higher spending (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.[24] 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 5 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. 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,

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

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.

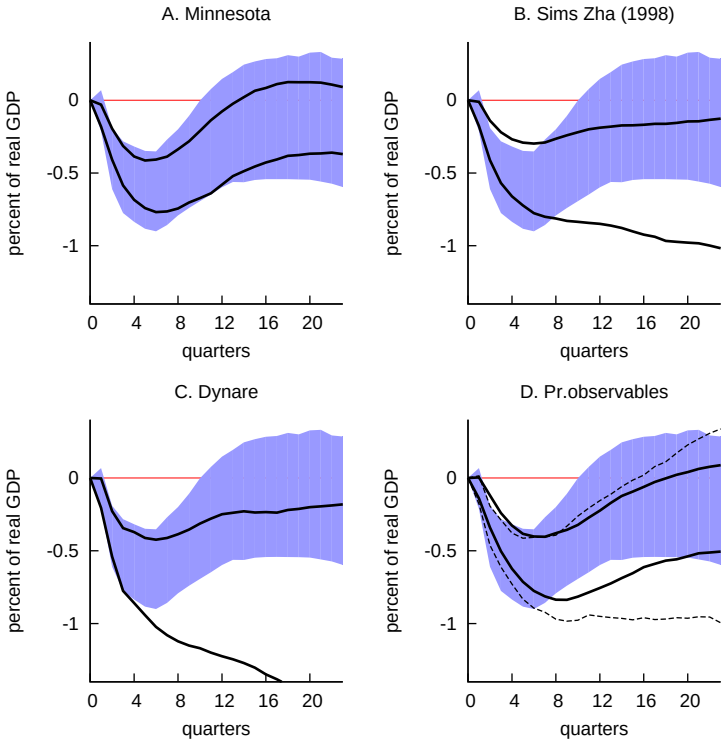

Figure 5 – Response of output to a monetary policy shock: OLS estimation (shaded area, the same across columns) and Bayesian estimations using four informative priors. Quantiles 0.05 and 0.95.

Figure 6, 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 Bayesian grounds.

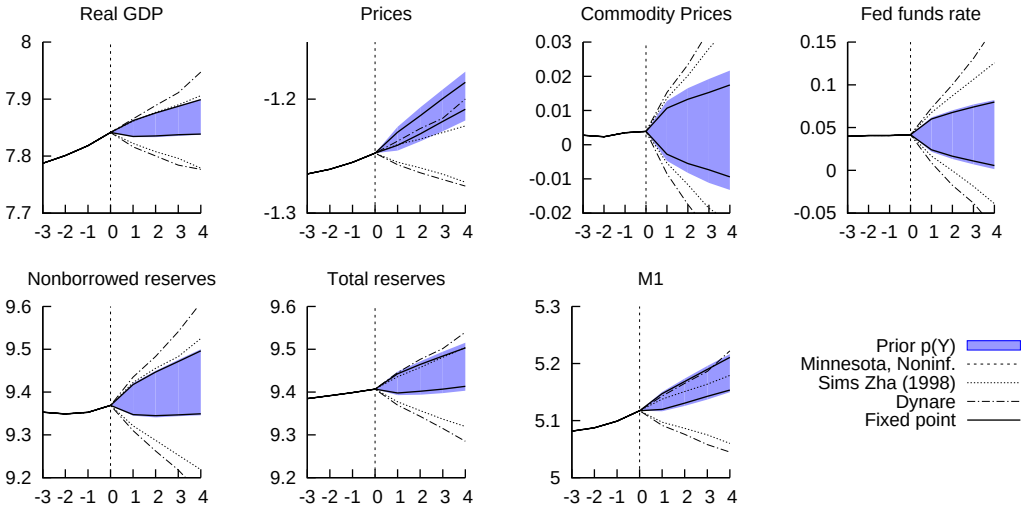

Figure 6 – Density of the observables implied by alternative priors. Quantiles 0.05 and 0.95 of the distribution in periods 1 to 4 of the estimation sample.

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.

We specify our prior on the growth rates in the initial P periods conditional on the observed pre-sample values $y^o_{-P+1}, ..., y^o_0$ (where the superscript ^{o} stands for 'observed'). In this way, the prior is akin to the assumptions in the so-called 'exact likelihood' approach, and to most frequentist small sample bias corrections that relate initial conditions to the true parameters, so it has the advantage of allowing to

compare the results with this literature.[25]

Thus, we specify a density on a $P \times N$ dimensional vector of observables $p_{\Delta y_1,...,\Delta y_P|y^o_{-P+1},...y^o_0}$ as our prior about observables. Specifying a prior on growth rates does not mean we impose a unit root, it is done only for convenience, obviously this prior 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.[26] 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 6, 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.

---

[25] See Jarociński and Marcet (2010), section 2 for a discussion.

[26] 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.

4.2.3 Results with the prior about observables

After specifying our density of the observables we again run the approximate conjugate algorithm from section 3.4. This time, using different random starting points g^0 for the algorithm we find different approximate fixed points consistent with the stated prior density of the observables. This happens because our prior about observables does not define a unique prior about parameters. Our prior states a distribution for a vector of Y's of dimension $NP = 28$, while the number of parameters for which it defines a prior p_θ is much larger.[27] Therefore, we need to impose some more restrictions in order to choose from among the many fixed points that we find. First, we restrict the marginal prior density of Σ to be the same as in the Minnesota, Sims Zha (1998) and Dynare priors. Then we find 300 approximate fixed points that satisfy the restriction on $p(\Sigma)$. We stop at 300 because the lessons drawn based on 300 fixed points 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 two: the one with the highest marginal likelihood and the one with the highest entropy. These choices somehow represent two opposite criteria: the highest marginal likelihood is the fixed point that best fits the data actually observed,[28] while maximum entropy can be interpreted as imposing as little prior knowledge as possible.[29] It also happens to be the case that the maximum-marginal-likelihood fixed point has one of the lowest entropies, and that

[27] B contains $N(NP+1)$ parameters and Σ contains $N(N+1)/2$ parameters. Since $N = 7$ and $P = 4$, the total dimension of the parameter vector is 231.

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

[29] 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).

the maximum-entropy fixed point has one of the lowest marginal likelihoods in the studied set of fixed points.

To check accuracy we look at the implications for observables of the approximate fixed points that we find. The solid lines in Figure 6 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.[30]

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 5. 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%).[31] More 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

[30]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 $p(\Sigma)$.

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

than with a flat prior, implying that we have incorporated useful information in the estimation.

The dashed line in panel D of Figure 5 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.

4.3 Romer and Romer (2004) VAR

In this subsection we estimate the effects of monetary policy shocks in the US following Romer and Romer (2004). They first construct monetary policy shocks using a version of the 'narrative approach.' Next, to measure the effects of the shocks on the economy they estimate a VAR with the log of industrial production, the log of the producer price index, and the cumulated monetary policy shocks. The observations are monthly and the estimation sample is from January 1966 to December 1996.[32] The VAR includes 36 lags.

[32]We took the data from the Data Appendix at the AER website.

4.3.1 Results with standard priors

Figure 7 shows the effects of a one percentage point monetary policy shock estimated with the standard priors. We plot the impulse responses for 48 quarters. 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 Romer and Romer.[33] The solid lines report the posteriors obtained with informative priors. The first three columns are for the standard priors: the Minnesota prior, the Sims and Zha (1998) prior and the Dynare prior.

Figure 7 shows that in this case the impulse responses from all four standard priors on parameters are rather similar to each other.

Figure 8, analogous to Figures 3 and 6, plots the quantiles for the distribution of the first few dates of observables implied by the priors. We can see that standard priors for VARs imply that very large output and price changes are highly likely. As in the previous examples, the large divergence of some of these priors questions from reasonable priors on observables questions the Bayesian grounds of the estimates derived from these priors.

4.3.2 A prior about observables

In most of the paper and in the previous two applied examples our reasoning has been that analysts have priors on observables based on their own experience and that this

[33]Romer and Romer (2004) include their cumulated shock as the last variable and compute the responses to the Choleski shock of this variable. This implies that the immediate responses of output and prices are zero by construction. For reasons we explain later we include the cumulated shock as the first variable and compute the responses to the Choleski shock. Hence, our responses show some small immediate effects of monetary shocks on output and prices. Otherwise, the impulse responses we obtain are similar to those they report, although they do imply somewhat smaller medium run effects of monetary policy.

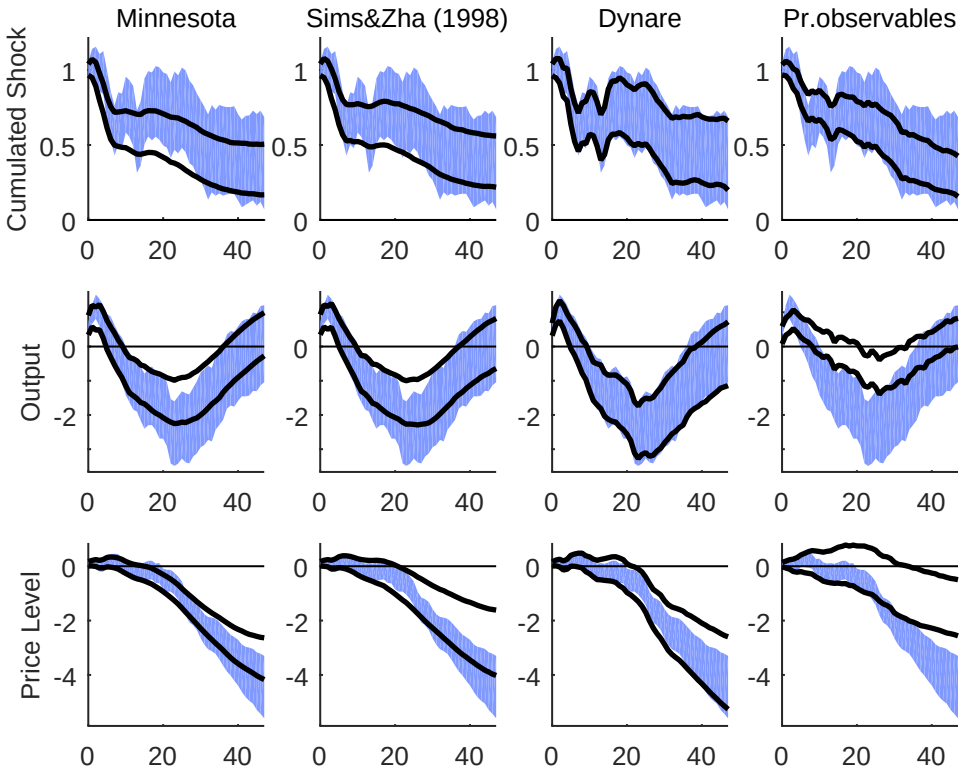

Figure 7 – Response to the Romer and Romer (2004) monetary shock: OLS estimation (shaded area, the same across columns) and Bayesian estimations using four informative priors. Quantiles 0.16 and 0.85.

should be used in estimation. In this example we take a somewhat different approach: we take take for granted that an economist does believe in a certain structural model and that this economist uses this model to express his prior knowledge about the behavior of the observables. A related motivation underlies the priors used by Del Negro and Schorfheide (2004) and others.

We now formulate a prior about observables, p_Y, based on a simple structural

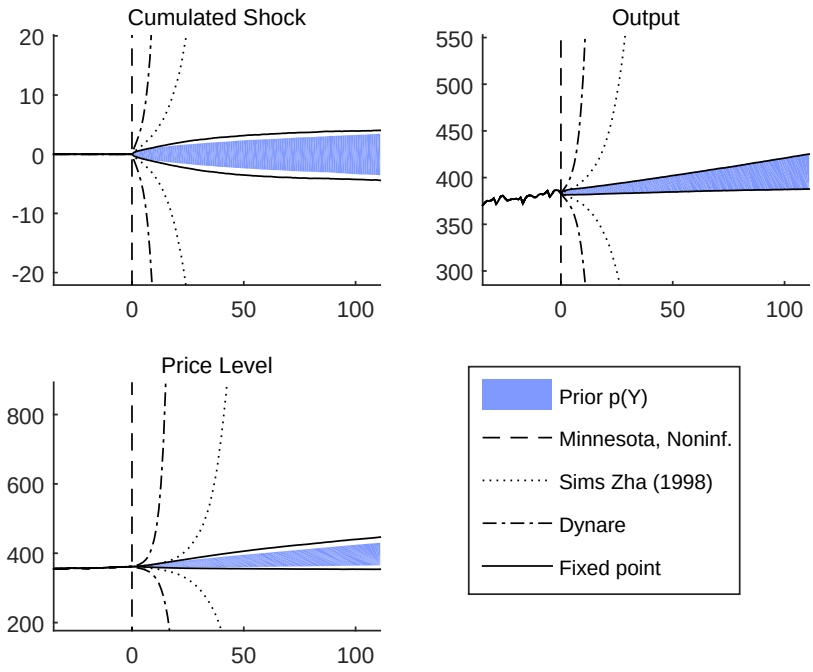

Figure 8 – Density of the observables implied by alternative priors. Quantiles 0.16 and 0.84 of the distribution in periods 1 to 111 of the estimation sample.

economic model:

$$y_t = b_1(r_t - E_t\pi_{t+1}) + E_t y_{t+1} + v_t, \tag{16a}$$

$$\pi_t = (1-\phi)E_t\pi_{t+1} + \phi\pi_{t-1} + \alpha(y_t - \bar{y}_t) + u_t, \tag{16b}$$

$$r_t = (1-\mu_3)\left[r + \pi_t + \mu_1(\pi_t - \pi^*) + \mu_2(y_t - \bar{y}_t)\right] + \mu_3 r_{t-1} + e_t, \tag{16c}$$

where y_t is the log of output, π_t is inflation, r_t is the nominal interest rate, and v_t, u_t and e_t are exogenous shocks. Equation (16a) is a New-Keynesian IS equation, (16b) is a New-Keynesian Phillips curve and (16c) is a Taylor-type monetary policy rule, with e_t representing a monetary policy shock. We take this model from McCallum (2001). McCallum discusses further details of the model and how it can be derived from optimizing behavior of households and firms under nominal rigidities. We chose this specific model because, first, we wanted a simple model in which monetary policy is nonneutral, to justify the interest in the effects of monetary policy shocks. Second,

we wanted a model with enough built-in inertia to produce delayed, hump-shaped impulse responses, which seem to be a feature of the data. However, although the model impulse responses are hump-shaped, monetary policy does have some immediate effects in it, and this is why we also allow for immediate effects of monetary policy in the VAR, by ordering the shock as the last variable.

The calibration of the model parameters is also taken from McCallum, except for adjustments that roughly adapt his quarterly calibration to our monthly data: we reduce α and μ_2 by a factor of 3, we increase μ_3 by raising it to the power 1/3, we reduce the variances of shocks and add autocorrelation of 0.5 to the shock processes, which are originally i.i.d. We divide shock standard deviations by 3, except for the monetary policy shock e_t, which has standard deviation of 0.0034 (the standard deviation of the shock constructed by Romer and Romer). When simulating the model, we add a trend output growth drawn from a normal distribution with mean 2.5 percent annualized and standard deviation of two percentage points. To inflation we add a trend inflation drawn from a normal distribution with mean 4 percent annualized and standard deviation of two percentage points. We also tried introducing uncertainty about shock standard deviations and some other model parameters, but these modifications did not affect our key findings.

To generate the density of the observables we repeat the following sequence of steps. 1) We draw all the parameters that have nondegenerate prior distributions. 2) We solve the model using the gensys procedure of Sims (2001). 3) We draw the shocks for 111 periods and compute the evolution of y_t and π_t implied by these shocks. 4) We compute the cumulated monetary policy shock e_t, the log level of output and the log level of prices. The obtained 111 periods long sequences of the three series are a draw from our prior density of the observables. We impose this density for 111 months, as then the dimension of the prior density of the observables (111×3) equals the

dimension of the prior density of the parameters B and Σ ($N(NP+1)+N(N+1)/2 = 333$). We use our approximate conjugate fixed point algorithm to translate this prior about observables into a prior about parameters. Starting from random points the algorithm converges to approximately the same density in less than 200 iterations, which takes about 30 minutes.

Figure 8, shows that McCallum's model implies reasonable evolution of the series and that the fixed point we compute approximates this prior on observables reasonably well.

4.3.3 Results with the prior about observables

The fourth column of Figure 7 shows that a researcher approaching Romer and Romer's data with McCallum's prior finds much weaker effects of monetary policy. The medium term responses of output and prices are reduced by a factor of three. This is perhaps unsurprising, as the impulse responses to monetary policy shocks found by Romer and Romer (2004) have quite a different dynamics from the impulse responses in McCallum (2001): in Romer and Romer prices only start falling after about 20 months, and before that time they remain basically constant, while in McCallum's model most of the response of prices occurs within the first 20 months. Therefore, a researcher who needs to reconcile this conflicting prior and data evidence settles for weak effects of monetary policy. This exercise highlights that the monetary policy effects in Romer and Romer are quite at odds with quantitative predictions of a standard New Keynesian model. It shows how a Bayesian researcher starting with McCallum's model should downweigh Romer and Romer's evidence.

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, we give a numerical algorithm to find this fixed point and we show 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. Often these standard priors 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 three 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.

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. Extending our analytical results would be useful. For example, our convergence result in Proposition 5 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. The algorithm can be used for non-parametric estimation along the lines discussed in footnote 12.

Appendix A Proofs of Propositions 1 to 5 and Result 1

Proof of Proposition 1

Clearly 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 at $g = p_\theta$.

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

Let $p_\theta > 0$ be the solution of (4) considered in the statement of the proposition and consider $\widetilde{p}_\theta$ another solution of (4) and. We have

$$E\left(\frac{\widetilde{p}_\theta(\theta)}{p_\theta(\theta)}\middle| Y\right) = \int_\Theta \frac{p_{Y|\theta}(Y;\overline{\theta})\ \widetilde{p}_\theta(\overline{\theta})}{\int_\Theta p_{Y|\theta}(Y;\cdot)\ p_\theta}\ d\overline{\theta} = \int_\Theta \frac{p_{Y|\theta}(Y;\overline{\theta})\ \widetilde{p}_\theta(\overline{\theta})}{\int_\Theta p_{Y|\theta}(Y;\cdot)\ \widetilde{p}_\theta}\ d\overline{\theta} = 1.$$

The first equality follows by writing $p_{\theta|Y}$ in terms of Bayes' formula, the second because $\widetilde{p}_\theta$ satisfies (4).

Take $\delta(\theta) = \frac{\widetilde{p}_\theta(\theta)}{p_\theta(\theta)} - 1$, completeness with respect to θ implies $\widetilde{p}_\theta = p_\theta$, therefore the solution is unique.■

Proof of Proposition 3

Consider the set $Y^0 \equiv \left\{Y \in \mathcal{Y} : p_{Y|\theta}(Y;\cdot) = 0\right\}$. Let $\mathbf{I}_{Y^o}$ be the indicator function. By definition of Y^0 we have that $E(I_{Y^o}(Y) \mid \theta) = 0$ for all θ. By completeness this implies that $Prob(\ Y \in\ Y^0) = 0$. Therefore $g^* > 0$ implies $\int_\Theta p_{Y|\theta}(Y;\cdot)\ g^* > 0$ a.s. in Y so that $\mathcal{F}$ is well defined at g^*.

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

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

Therefore, taking $\delta(Y) = \frac{p_Y(Y)}{\int_\Theta p_{Y|\theta}(Y;\cdot)\ g^*} - 1$, completeness implies that $\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g^* = p_Y(\overline{Y})$ for almost all $\overline{Y} \in \mathcal{Y}$. ■

Proof of Proposition 4

We first show that $\mathcal{F}(g^*)$ is well defined. Since $\pi_{kj} \geq 0$ and $g_k > 0$ we have

$$\sum_k \pi_{kj} g_k^* \geq 0 \text{ all } j = 1, ..., N. \tag{A.1}$$

Since $g_i > 0$ for all i, the only way that (A.1) could hold as equality for some given j is if $\pi_{kj} = 0$ for all k. But this would violate invertibility of Π. Therefore $\sum_k \pi_{kj} g_k > 0$ for all j and $\mathcal{F}(g)$ is well defined.

Using $g_i^* > 0$, the fixed point condition implies that

$$\sum_j \frac{\pi_{ij}}{\sum_k \pi_{kj} g_k^*}\ p_Y(\overline{Y}_j) = 1 \quad \text{for all } i = 1, ..., N. \tag{A.2}$$

Let $h \in R^N$ have $h_j = \frac{p_Y(\overline{Y}_j)}{\sum_k \pi_{kj} g_k^*}$ as typical element. Let $\mathbf{1} \in R^N$ have all elements equal to 1. Equation (A.2) can be written as

$$\Pi h = \mathbf{1} \tag{A.3}$$

Since all the rows of Π add up to 1 we have $\Pi\mathbf{1} = \mathbf{1}$. Premultiplying both sides of the last equation by Π^{-1} we have that $h = \mathbf{1}$ and it follows that

$$\sum_k \pi_{kj} g_k^* = p_Y(\overline{Y}_j) \quad \text{for all } j = 1, ..., N \tag{A.4}$$

so that g^* solves (3). ■

Proof of Proposition 5

At the beginning of the proof of Proposition 4 we argued that if $g^* > 0$ then $\Pi g^* > 0$. The same argument proves that under assumptions *i)* and *ii)* in the current proposition $\Pi g_\theta > 0$. Therefore $\mathcal{F}(g)$ is well defined near g_θ and taking derivatives of $\mathcal{F}$ mechanically we have

$$\frac{\partial \mathcal{F}(g)_i}{\partial g_n} = \begin{cases} \sum_j \frac{\pi_{ij}}{\sum_k \pi_{kj} g_k} \, p_Y(\overline{Y}_j) - \sum_j \frac{\pi_{nj} \, \pi_{ij} \, p_Y(\overline{Y}_j)}{\left(\sum_k \pi_{kj} g_k\right)^2} g_i & \text{for } n = i \\ -\sum_j \frac{\pi_{nj} \, \pi_{ij} \, p_Y(\overline{Y}_j)}{\left(\sum_k \pi_{kj} g_k\right)^2} g_i & \text{for } n \neq i. \end{cases}$$

Using $\sum_k \pi_{kj} g_{\theta,k} = p_Y(\overline{Y}_j) > 0$, letting Δ^* be the matrix with a typical element $\Delta^*_{in} = \sum_j \pi_{nj} \frac{\pi_{ij} g_{\theta,i}}{\sum_k \pi_{kj} g_{\theta,k}}$, we can evaluate this derivative at g_θ to obtain

$$\frac{\partial \mathcal{F}(g_\theta)_i}{\partial g_n} = \begin{cases} 1 - \Delta^*_{in} & \text{for } n = i \\ -\Delta^*_{in} & \text{for } n \neq i, \end{cases}$$

so that

$$\frac{\partial \mathcal{F}(g_\theta)}{\partial g'} = I - \Delta^*. \tag{A.5}$$

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

$$\lambda_n \text{ is a real number and } 0 < \lambda_n \leq 1 \tag{A.6}$$

It is easy to verify that the rows of $\Delta^{*\prime}$ add up to 1. A well known property of such matrices is that $|\lambda_n| \leq 1$ for all $n = 1, ..., N$.

Next we discard the possibility that the eigenvalues λ_n are complex and/or negative. Let G^* and $\mathcal{D}$ be diagonal matrices with the j-th diagonal entry equal to $g_{\theta,j}$ and $\frac{1}{\sum_k \pi_{kj} g_{\theta,k}}$ respectively. We can write

$$G^* \Delta^* = G^* \Pi \mathcal{D} \Pi' G^* \tag{A.7}$$

showing that $G^* \Delta^*$ is a symmetric positive semidefinite matrix. Furthermore, since g_θ and $\sum_k \pi_{kj} g_{\theta,k}$ are strictly positive and Π is invertible all matrices involved in

the right side of (A.7) are invertible so that no eigenvalues of $G^*\Delta^*$ can be zero. Therefore, $G^*\Delta^*$ is positive definite, hence all its eigenvalues are real and strictly positive. It remains to show that all eigenvalues of Δ^* inherit this property.

Obviously

$$\Delta^* = (G^*)^{-1} G^* \Delta^*. \tag{A.8}$$

Clearly $(G^*)^{-1}$ is symmetric and positive definite and we already know that $G^*\Delta^*$ is symmetric and positive definite. When two matrices are symmetric and positive definite then all the eigenvalues of their product are real and strictly positive (e.g. this is a special case of Serre (2010) Proposition 6.1). Hence, we have shown that all real numbers $\lambda_n > 0$ for all n. This ends the proof of (A.6).

The eigenvalues of $(I - \Delta^*)$ are $1 - \lambda_n$, hence by (A.6) and (A.5) we have that all eigenvalues of $\frac{\partial \mathcal{F}(g_\theta)}{\partial g'}$ are strictly less than one in absolute value. A standard argument implies that successive approximations on $\mathcal{F}$ locally converge to g_θ. ■

Proof of Result 1

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

$$= \int_\mathcal{Y} \left(\int_\Theta q(\overline{\theta}) \; p_g(\overline{\theta}|\cdot) \; d\overline{\theta} \right) \; p_Y = E_{p_Y} \left(\; E_{p^g(\cdot|Y)}(q(\theta)) \; \right) \tag{A.9}$$

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}. This proves (9). ■

Appendix B A Discrete Approximation to the continuous case

We describe a discretization of continuous distributions and find conditions guaranteeing that the fixed points of this modified problem converge to a solution of the

continuous inverse equation (4) as the step size goes to zero. Combining this result with Proposition 5 we can state that for sufficiently many iterations on $\mathcal{F}$ and sufficiently small step size ε we can approximate the continuous p_θ that solves (4) arbitrarily well.

Building $\varepsilon-$partitions

Fix a scalar $\varepsilon > 0$. An $\varepsilon-$partition is a collection of non-overlapping intervals $\{\mathbf{Y}_i^\varepsilon\}_{i=1}^{N_\varepsilon}$ where $\mathbf{Y}_i^\varepsilon \subset \mathcal{Y} \subset \mathcal{R}^M$ with $N_\varepsilon < \infty$ (that cover the support of Y. Formally, we require that $\mathbf{Y}_i^\varepsilon \cap \mathbf{Y}_j^\varepsilon = \varnothing$ for all $i \neq j$ and that $\cup_{i=1}^{N_\varepsilon} \mathbf{Y}_i^\varepsilon = \text{supp}(\mathcal{Y})$ where $\text{supp}(\mathcal{Y})$ denotes the set of Y values that have a positive density for some $\theta \in \Theta$. The sides of all intervals are either of length less than ε or infinite. If $\mathcal{Y}$ is not compact we allow for infinite intervals but the probability of sets $\mathbf{Y}_i^\varepsilon$ with infinite sides has to go to zero as $\varepsilon \to 0$.

More specifically, these intervals can be constructed as follows: for each dimension $m = 1, ..., M$ we choose a given set of $I_\varepsilon < \infty$ interval endpoints $Y_m^{\varepsilon,i}$, $i = 1, ..., I_\varepsilon$ where $Y_m^{\varepsilon,i} \in R$ for $i = 2, ..., I_\varepsilon - 1$ but $Y_m^{\varepsilon,1}, Y_m^{\varepsilon,I_\varepsilon} \in R \cup \{-\infty, \infty\}$. The endpoints have to cover the whole support so that $Y_m^{\varepsilon,1} \leq \inf_{\text{supp}(\mathcal{Y}_m)}$ and $Y_m^{\varepsilon,I_\varepsilon} \geq \sup_{\text{supp}(\mathcal{Y}_m)}$ where $\mathcal{Y}_m$ is the projection of $\text{supp}(\mathcal{Y})$ on its m-th coordinate. We require $Y_m^{\varepsilon,i} < Y_m^{\varepsilon,i+1}$ $i = 1, ..., I_\varepsilon - 1$, $|Y_m^{\varepsilon,i} - Y_m^{\varepsilon,i+1}| < \varepsilon$ for $i = 2, ..., I_\varepsilon - 2$ and for the lowest endpoint $|Y_m^{\varepsilon,1} - Y_m^{\varepsilon,2}| < \varepsilon$ if $\inf_{\text{supp}(\mathcal{Y}_m)} > -\infty$, similarly for the highest endpoint $Y_m^{\varepsilon,I_\varepsilon}$. Finally, in the case $\inf_{\text{supp}(\mathcal{Y}_m)} = -\infty$ (sup) we require that $Y_m^{\varepsilon,2} \to -\infty$ ($Y_m^{\varepsilon,I_\varepsilon - 1} \to \infty$).

We consider all intervals of the form $\prod_{m=1}^{M} (Y_m^{\varepsilon,i_m}, Y_m^{\varepsilon,i_m+1}]$ for some $i_m \in \{1, ..., I_\varepsilon - 1\}$, clearly $\mathcal{Y}$ is included in the union of these intervals. Finally we construct sets $\mathbf{Y}_i^\varepsilon \subset \mathcal{Y}$ by overlapping each interval with $\mathcal{Y}$, that is we set $\mathbf{Y}_i^\varepsilon = \text{supp}(\mathcal{Y}) \cap \prod_{m=1}^{M} (Y_m^{\varepsilon,i_m}, Y_m^{\varepsilon,i_m+1}]$ for all the intervals where the intersection is non-empty (empty sets have to be excluded if Π^ε is to be invertible). Let $N_\varepsilon \leq (I_\varepsilon)^M$ be the number of these intervals.

We consider analogous partitions $\{\boldsymbol{\theta}_i^\varepsilon\}_{i=1}^{N_\varepsilon}$ of Θ, where the number of sets N_ε is the same both in the partitions of $\mathcal{Y}$ and Θ. However, for our proof to work we need to exclude intervals for θ with infinite sides, so that all the endpoints $\theta_m^{\varepsilon,i}$, $i = 1, ..., I_\varepsilon$ are such that $|\theta_m^{\varepsilon,i}| < \infty$. In the case where Θ has infinite support we require $-\theta_m^{\varepsilon,1}, \theta_m^{\varepsilon,I_\varepsilon} \to \infty$ as $\varepsilon \to 0$. This guarantees that all $\boldsymbol{\theta}_i^\varepsilon$ are compact and $\cup_{i=1}^{N_\varepsilon} \boldsymbol{\theta}_i^\varepsilon \nearrow \text{supp}(\Theta)$ as $\varepsilon \to 0$.

Discretizing p_Y and $p_{Y|\theta}$

We discretize p_Y by forming an N_ε-dimensional probability vector as follows. Let

$$p_{Y,i}^\varepsilon \equiv \int_{\mathbf{Y}_i^\varepsilon} p_Y$$

and let p_Y^ε be the vector with a typical element $p_{Y,i}^\varepsilon$. Clearly p_Y^ε defines a discrete probability distribution of Y.

We discretize $p_{Y|\theta}$ defining

$$\pi_{ij}^\varepsilon \equiv \int\limits_{\mathbf{Y}_j^\varepsilon \times \boldsymbol{\theta}_i^\varepsilon} p_{Y|\theta}$$

and letting Π^ε be the matrix with a typical element π_{ij}^ε. Clearly Π^ε is a special case of the likelihood matrix Π considered in section 3.1, as its rows add up to 1. This follows from the fact that the ε-partition is chosen so that $\cup_{i=1}^{N_\varepsilon} \mathbf{Y}_i^\varepsilon = \text{supp}(\mathcal{Y})$.

Let $g_\theta^\varepsilon \in R^{N_\varepsilon}$ be a discrete distribution that satisfies the discrete inverse equation

$$\Pi^{\varepsilon\prime} g_\theta^\varepsilon = p_Y^\varepsilon \tag{B.1}$$

We assume for now that this solution g_θ^ε exists. Let G_θ^ε be a cumulative distribution function for a continuous random variable θ defined as being uniform in $\boldsymbol{\theta}_j^\varepsilon$ and such that $\int\limits_{\boldsymbol{\theta}_j^\varepsilon} dG_\theta^\varepsilon = g_{\theta,j}^\varepsilon$ for all $j = 1, ..., N_\varepsilon$. Notice that G_θ^ε is well defined because we have restricted the intervals $\boldsymbol{\theta}_j^\varepsilon$ to be compact, a uniform distribution would not exist over an interval with an infinite side.

We prove that G_θ^ε becomes arbitrarily close to a solution of the continuous inverse equation (4) as $\varepsilon \to 0$. We first prove the following Lemma.

Lemma 1. *Fix $\varepsilon-$partitions of $\mathcal{Y}$ and Θ. We make the following assumptions on the likelihood function $p_{Y|\theta}$ and the density of observables p_Y.*

i) Π^ε is invertible for all ε.

ii) $p_{Y|\theta}$ is bounded, $p_{Y|\theta}(\overline{Y};\cdot)$ is continuous a.s. in $\overline{Y}$ with respect to p_Y and p_Y is continuous in $\mathcal{Y}$.

iii) The solution to (B.1) satisfies $g_\theta^\varepsilon \geq 0$.

Then the limit of any convergent subsequence of $G_\theta^{\varepsilon_k}$ solves (4). More precisely, for a subsequence $\{G_\theta^{\varepsilon_k}\}_{k=1}^\infty$ with $\varepsilon_k \to 0$ such that

$$G_\theta^{\varepsilon_k} \to \widetilde{G}_\theta \text{ weakly as } k \to \infty$$

for some distribution $\widetilde{G}_\theta$, we have that $\widetilde{G}_\theta$ solves (4).

Invertibility of Π^ε can be checked numerically for a given ε. The interpretation of this assumption is similar to the interpretation of completeness: the model should identify θ for any possible value of the observables.

Proof of Lemma 1

Given $\overline{Y} \in \mathcal{Y}$ it follows from the assumptions that $\int_{-\infty}^{\overline{Y}} p_{Y|\theta}(\widetilde{Y};\cdot)d\widetilde{Y}$ is a bounded continuous function of θ, therefore by weak convergence

$$\int_\Theta \left[\int_{-\infty}^{\overline{Y}} p_{Y|\theta}(\widetilde{Y};\cdot)d\widetilde{Y}\right] dG_\theta^{\varepsilon_k} \to \int_\Theta \left[\int_{-\infty}^{\overline{Y}} p_{Y|\theta}(\widetilde{Y};\cdot)d\widetilde{Y}\right] dG_\theta \text{ as } k \to \infty. \tag{B.2}$$

Applying Fubini's theorem to both sides of this limit we have

$$\int_{-\infty}^{\overline{Y}} \left[\int_\Theta p_{Y|\theta}(\widetilde{Y};\cdot)dG_\theta^{\varepsilon_k}\right] d\widetilde{Y} \to \int_{-\infty}^{\overline{Y}} \left[\int_\Theta p_{Y|\theta}(\widetilde{Y};\cdot)dG_\theta\right] d\widetilde{Y} \text{ as } k \to \infty. \tag{B.3}$$

For a given k and subset $\mathbf{Y}_j^{\varepsilon_k}$

$$\int_{\mathbf{Y}_j^{\varepsilon_k}} \left[\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)dG_\theta^{\varepsilon_k}\right] d\overline{Y} = \int_{\mathbf{Y}_j^{\varepsilon_k}} \left[\sum_{i=1}^{N_{\varepsilon_k}} \int_{\theta_i^{\varepsilon_k}} p_{Y|\theta}(\overline{Y};\cdot)\ g_i^{\varepsilon_k}\right] d\overline{Y} =$$

$$\sum_{i=1}^{N_{\varepsilon_k}} \int_{\mathbf{Y}_j^{\varepsilon}\times\theta_i^{\varepsilon}} p_{Y|\theta}(\overline{Y};\overline{\theta})\ g_i^{\varepsilon_k}\ d(\overline{Y},\overline{\theta}) = \sum_{i=1}^{N_{\varepsilon_k}} \pi_{ij}^{\varepsilon_k} g_i^{\varepsilon_k} = p_{Y,j}^{\varepsilon} \tag{B.4}$$

where the first equality follows from the fact that $\boldsymbol{\theta}_i^{\varepsilon}$ are non-overlapping, that $G_{\theta}^{\varepsilon_k}$ puts probability one on $\cup_{i=1}^{N_{\varepsilon}} \boldsymbol{\theta}_i^{\varepsilon}$ and that $G_B^{\varepsilon_k}$ is uniform in each subset $\boldsymbol{\theta}_i^{\varepsilon}$, the third equality follows from the definition of $\pi_{ij}^{\varepsilon_k}$ and the last from (B.1).

Let $\left\{i : \mathbf{Y}_i^{\varepsilon_k} \subset (-\infty, \overline{Y}]\right\}$ include the indexes of all the sets in the ε_k-partition that are fully included in the interval $(-\infty, \overline{Y}]$. We have

$$\int_{\cup\left\{\mathbf{Y}_i^{\varepsilon_k}:\mathbf{Y}_i^{\varepsilon_k}\subset(-\infty,\overline{Y}],i=1,...,N_{\varepsilon_k}\right\}} \left[\int_{\Theta} p_{Y|\theta}(\overline{Y};\cdot) dG_{\theta}^{\varepsilon_k}\right] d\overline{Y} = \sum_{i:\mathbf{Y}_i^{\varepsilon_k}\subset(-\infty,\overline{Y}]} p_{Y,i}^{\varepsilon_k} \to \int_{-\infty}^{\overline{Y}} p_Y \text{ as } k \to \infty. \tag{B.5}$$

The equality follows from the fact that the intervals $\mathbf{Y}_i^{\varepsilon_k}$ are disjoint and (B.1). One has to be careful arguing for the convergence part in (B.5), one can not simply claim that the set $\cup\left\{\mathbf{Y}_i^{\varepsilon_k} : \mathbf{Y}_i^{\varepsilon_k} \subset (-\infty, \overline{Y}], i = 1, ..., N_{\varepsilon_k}\right\}$ converges to $(-\infty, \overline{Y}]$,since convergence of sets is a problematic concept. Convergence in (B.5) follows from the following argument. Let the m-th element of $Y^{\varepsilon}(\overline{Y}) \in \mathcal{R}^M$ be defined as the highest interval endpoint in the $\varepsilon-$partition that is lower than $\overline{Y}$, more precisely,

$$Y^{\varepsilon}(\overline{Y})_m = \max_{Y_m^{\varepsilon,i} \leq \overline{Y}_m} \left\{Y_m^{\varepsilon,i}; i = 1, ..., I^{\varepsilon}\right\}$$

Then we have

$$\left|\sum_{i:\mathbf{Y}_i^{\varepsilon_k}\subset(-\infty,\overline{Y}]} p_{Y,i}^{\varepsilon_k} - \int_{-\infty}^{\overline{Y}} p_Y\right| = \left|\int_{-\infty}^{Y^{\varepsilon_k}(\overline{Y})} p_Y - \int_{-\infty}^{\overline{Y}} p_Y\right| = \left|\int_{Y^{\varepsilon_k}(\overline{Y})}^{\overline{Y}} p_Y\right|$$

By construction $\left|Y^{\varepsilon}(\overline{Y})_m - \overline{Y}_m\right| < \varepsilon$ hence the sets $\left\{Y \in \mathcal{R}^M :\ Y^{\varepsilon}(\overline{Y})_m \leq Y_m \leq \overline{Y}_m\right\}$ have Lebesgue measure that converges to zero, therefore $\left|\int_{Y^{\varepsilon_k}(\overline{Y})}^{\overline{Y}} p_Y\right| \to 0$ because of continuity of p_Y. The convergence part in (B.5) follows.

A similar argument gives

$$\int\limits_{\cup\left\{\mathbf{Y}_i^{\varepsilon_k}:\mathbf{Y}_i^{\varepsilon_k}\subset(-\infty,\overline{Y}],i=1,\ldots,N_{\varepsilon_k}\right\}}\left[\int\limits_{\Theta}p_{Y|\theta}(\overline{Y};\cdot)dG_\theta^{\varepsilon_k}\right]d\overline{Y}\to\int\limits_{-\infty}^{\overline{Y}}\left[\int\limits_{\Theta}p_{Y|\theta}(\overline{Y};\cdot)dG_\theta\right]d\overline{Y}$$

and by (B.3) we have

$$\int\limits_{-\infty}^{\overline{Y}}\left[\int\limits_{\Theta}p_{Y|\theta}(\overline{Y};\cdot)dG_\theta\right]d\overline{Y}=\int\limits_{-\infty}^{\overline{Y}}p_Y,$$

implying that the inverse equation (4) holds for the distribution functions implied by the densities p_θ and p_Y.■

Assuming uniqueness we have

Proposition 6. ***(Approximation by step functions)*** *If the (continuous) inverse equation (4) has a unique solution density p_θ with a corresponding cdf G_θ, and the assumptions of Lemma 1 hold, then $G_\theta^\varepsilon\to G_\theta$ weakly as $\varepsilon\to 0$.*

The proof follows immediately from the previous lemma and the fact that the space of distributions is compact so that any sequence has a convergent subsequence.

Appendix C 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 (12)-(13). 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.

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.

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.

Giannone, D., Lenza, M., and Primiceri, G. E. (2016). Priors for the long run. unpublished.

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. (2016). An inflation-predicting measure of the output gap in the euro area. Working Paper Series 1966, European Central Bank.

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.

Martin, R. and Ghosh, J. K. (2008). Stochastic approximation and Newton's estimate of a mixing distribution. *Statistical Science*, 23(3):365–382.

McCallum, B. T. (2001). Should monetary policy respond strongly to output gaps? *American Economic Review Papers and Proceedings*, 91(2):258–262.

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.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: derivation and implications. *The American Economic Review*, 94(4):1055–1084.

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.

Rutanen, K. (2011). Tim matlab 1.2.0. Matlab toolbox.

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

Sims, C. (2001). Solving linear rational expectations models. *Journal of Computational Economics*, 20(1-2):1–20.

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.

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

Wang, Q., Kulkarni, S. R., and Verdu, S. (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Information Theory*, 55(5):1961–1975.

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