Priors about Observables in Vector Autoregressions[*]

Marek Jarociński

European Central Bank

Albert Marcet

Institut d'Anàlisi Econòmica CSIC, ICREA, Barcelona GSE,

UAB, Bank of Spain Professor

March 18, 2013

Abstract

Standard practice in Bayesian VARs is to formulate priors on the autoregressive parameters, but economists and policy makers actually have priors about the behavior of observable variables. We show how this kind of prior can be used in a VAR 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 very large number of parameters. We prove various

[*] We thank Gianni Amisano, Manolo Arellano, Stéphane Bonhomme, Jean-Pierre Florens, Bartosz Maćkowiak, Leonardo Melosi, Thomas J. Sargent, Frank Schorfheide, Chris Sims and Jim Stock for useful conversations, and Harald Uhlig for his comments on the earlier version of this paper. 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.

convergence theorems for the algorithm. As an application, we first show that the results in Christiano et al. (1999) are very sensitive to the introduction of various priors that are widely used. These priors turn out to be associated with undesirable priors on observables. But an empirical prior on observables helps clarify the relevance of these estimates: we find much higher persistence of output responses to monetary policy shocks than the one reported in Christiano et al. (1999) and a significantly larger total effect.

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

JEL codes: C11, C22, C32

1 Introduction

Vector autoregressions (VARs) are frequently estimated using the Bayesian approach. Formulating convincing priors for VARs is crucial, because in practice the sample is often small and priors have a large effect on the results. Usual practice is to formulate priors directly on the autoregressive parameters, but it is difficult to reflect our prior knowledge about the economy in this way. In fact, economists hold prior ideas about the behavior of observed time series, not about parameters.

For example, perhaps an economist could formulate a prior that the growth rate of GDP in a certain period has a mean around 1.5 or 2 percent per annum and that the prior standard deviation around this mean is between, say, .8 and 1.5 percentage points. Translating this simple statement about observables into a joint prior on all VAR parameters is not easy because, in a time series model, a prior distribution on observables maps into a prior distribution on parameters via a complex inverse problem.

In this paper we show that priors on observables are useful in Bayesian VARs. We provide an algorithm to solve the implied inverse problem, we prove that this algorithm converges and that it works well in practice. We apply this framework to the VAR of Christiano et al. (1999).

Our proposal is to formulate explicitly a prior on *observable* time series. The implied prior distribution of parameters is the solution of an inverse problem, a Fredholm equation of the first kind. We reformulate this inverse problem as a fixed point of a certain mapping, and 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.

As an application we reexamine the study of monetary policy shocks in the U.S. in Christiano et al. (1999) (CEE). We find that this influential paper would show widely disparate results under four commonly used alternative priors. Two of these priors even contradict long-run neutrality of money. There is little guidance in the literature about how to choose among these priors, they are all standard, commonly used, and an empirical economist or a policy maker may be quite confused about the results.[1]

Since these standard priors are statements about uninterpretable VAR parameters it is unlikely that introspection will help us decide which prior better reflects our beliefs. In fact we find that these apparently reasonable 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 Bayesian point of view.

We propose incorporating "reasonable" prior knowledge on the observable as a resolution. We reestimate CEE under this light and we find a much more persistent

[1] The sensitivity of long-run effects of shocks in this application is not specific to the Bayesian approach. Classical small sample bias corrections have the same problem (see Jarociński and Marcet (2010), section 4.2).

effect of a monetary shock on output, though a weaker effect in the first two years, relative to CEE. We find long-run neutrality of money. Results are robust to reasonable changes in the prior. This example shows, first, that our prior on observables may be useful in clarifying empirical results. Second, it reduces posterior variance relative to the noninformative prior by incorporating useful information in the inference. Third, our algorithm works well in practice in a relatively large VAR where the fixed point we compute has hundreds of parameters.

Another advantage of our prior is that it produces good results when evaluated from a classical perspective. In Jarociński and Marcet (2010), section 5, 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 application to CEE. The appendix contains the proofs. An appendix available online provides additional 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.[2]

Popular priors for VARs, such as the Minnesota prior of Doan et al. (1984), dummy observation priors of Sims and Zha (1998) and others or the priors reflecting selected moments of the observables of Christiano et al. (2011) are often motivated by the implied behavior of the series, but the connection between prior on parameters and on observables is informal and/or indirect. Villani (2009) states a prior on the steady

[2] Kadane et al. (1996) is a small scale time series application.

state mean of stationary variables, but his approach can not be extended to other statements about observables. Our paper is the first to derive a VAR posterior from a prior on observables applying strict probability theory.

Inverse problems have recently attracted a lot of interest in microeconometrics, see Carrasco et al. (2007) for a survey. In this literature issues of consistency and asymptotic distribution are crucial, while we are interested in the computation of a fixed distribution. More importantly, the numerical methods used in this literature would be unfeasible for the high-dimensional problems that we face. To explain why, let us discuss the computational approach in two recent and influential papers within this literature.

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

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

One common theme in the literature just mentioned is whether or not a solution exists and the inverse problem is ill-posed. We do not focus on these issues in this paper. Our approximate fixed point gives an exact solution for a certain prior on observables and the analyst can check ex-post if this prior on observables captures approximately his prior. This alleviates the problem of existence. 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. For example, Carrasco and Florens (2011) use a Tikhonov regularization for the same purpose. More work to study 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, and these restrictions involve additional complications. Another advantage of our algorithm is that it obtains proper densities at each step 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., the current value of the estimated quantity adjusts with each new observation that is incorporated but, for simplicity, the implications of the new observation for the past estimated quantity are disregarded. On-line estimation was designed for practical applications when relevant information arrives very rapidly, faster than the new information can be processed optimally by a computer. 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 the last information is likely to be more important than, say, maximizing the likelihood function as each new piece of information arrives. But in academic papers using on-line estimators is less justified, it adds noise to the estimation and it can be quite inaccurate.[3] For example, one well-known side-effect of on-line estimation is

[3] On-line estimators have received some attention already in economics. For example, convergence

that Newton's estimates depend on the ordering of the observations.

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 applications. 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. We show 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.[4]

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 formulating a subjective prior p_θ directly. 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 and then apply Bayes' formula in a standard way.

An example

Let variable y follow a univariate AR(1) model

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

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

results for these algorithms have been fundamental for the literature on convergence of least squares learning models to rational expectations, as in Marcet and Sargent (1989) and Evans and Honkapohja (2002). But on-line estimators have been rarely used for computational purposes.

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

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 initial periods, for example[5]

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

for given $\mu_\Delta, \sigma_\Delta^2$. The researcher is not stating that y has a unit root, this is just her prior about Δy_1, it is compatible with many values for ρ.

To translate the prior on observables (ii) into the implied prior on α, ρ note that

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

and provided that $\sigma_\Delta^2 \geq \sigma_\epsilon^2$ the implied prior on α, ρ satisfies:

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

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, as would be the case for a prior variance on observables $\sigma_\Delta^2 < \sigma_\epsilon^2$. Second, there may be more than one solution, since (iii) only imposes a restriction on a linear combination of α, ρ. To obtain a proper prior on parameters we need to complement (iii) with an additional assumption, for example, about the marginal distribution of α or about the distribution of Δy_2. Third, equation (iii) 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, since specifying prior correlations between parameters is difficult, but imposing zero prior correlation on parameters often leads to unreasonable priors on observables.

[5] Normality and a fixed σ_ε^2 are convenient for an analytic solution in this example. The algorithm in section 3 does not need these assumptions, in fact we estimate σ_ε^2. It is also for convenience that the prior statement is only about the first observation $t = 1$, in general we use priors on more dates.

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

$$\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\; p_\theta = p_Y(\overline{Y}) \quad \text{for almost all } \overline{Y} \in \mathcal{Y} \tag{1}$$

where the "almost all" statement is with respect to p_Y. Our task is, given the known densities p_Y and $p_{Y|\theta}$, to find the prior density p_θ that satisfies the functional equation (1). 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 insure this in several ways by adjusting p_Y. Multiple solutions are likely to 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 for one approach to selecting one from the potentially multiple solutions.

3 Fixed point formulation

Fredholm equations like (1) can rarely be solved analytically.[6] We now reformulate our inverse problem in terms of a fixed point problem that facilitates computation. We present some results on necessity and sufficiency of the fixed point condition. We propose an algorithm for finding the fixed point by successive approximations and prove two convergence results for this algorithm. Finally, we describe the approximate

[6] The AR(1) example of section 2 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 for example, Jarociński and Marcet (2010) Appendix C for a detailed discussion.

conjugate fixed point iteration that we use in practice and we show how to check accuracy.

Let $g : \Theta \to \mathcal{R}_+$ be a probability density on Θ. Define the mapping $\mathcal{F}$:

$$\mathcal{F}(g)(\overline{\theta}) \equiv \int_{\mathcal{Y}} \frac{p_{Y|\theta}(\overline{Y};\overline{\theta})\ g(\overline{\theta})}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g}\ p_Y(\overline{Y})\ d\overline{Y} \quad \text{for all } \overline{\theta} \in \Theta. \tag{2}$$

Let us comment on the notation. First, we have written the integrals in (1) and in (2) in terms of densities, although in some places in the paper we will actually think of integrating over discrete probability distributions instead. This will be an obvious modification of (1) and (2) and to conserve space we do not write it explicitly now. Second, the mapping $\mathcal{F}$ is indexed by $p_{Y|\theta}$ and p_Y but we leave this dependence implicit to avoid notational clutter.

Clearly $\mathcal{F}(g)$ is itself a density.

$\mathcal{F}(g)$ 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 some density 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 $p_Y(\overline{Y})$. The mapping $\mathcal{F}$ is well defined whenever $\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g$ is non-zero almost surely in p_Y.

We now show that there is a close relation between solutions of (1) and fixed points of the mapping $\mathcal{F}$.

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

Uniqueness of solutions to (1) and sufficiency of a fixed point condition $\mathcal{F}(g^*) = g^*$ are closely related through the concept of completeness. We say that the joint distribution of random variables a and b, $p_{a,b}$, is complete with respect to a when it holds that if a measurable function $\delta : \mathcal{A} \to R$ (where $\mathcal{A}$ is the space of a) satisfies $E(\delta(a) \mid b) = 0$ for all $b \in \mathcal{B}$ (where $\mathcal{B}$ is the space of b) then $\delta = 0$ a.s. in $\mathcal{A}$.

Proposition 2. ***(Uniqueness)***. *Assume that $p_{\theta,Y}$ is complete with respect to θ, there exists a solution of (1) satisfying $p_\theta > 0$. Then the solution to (1) is unique.*

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 (1).*

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

These completeness conditions essentially mean that the model $p_{Y|\theta}$ is identified, in other words that values of Y a.s. carry relevant information about the value of θ and vice versa. The relationship between completeness and identification has been the object of much recent research in non-parametric estimation, starting with Newey and Powell (2003). Completeness may be hard to check, see for example the recent paper by Canay et al. (2012).

The above propositions suggest that instead of trying to solve problem (1) directly we can search for fixed points of the mapping $\mathcal{F}$. Let us state, for future reference, a simple algorithm for searching for fixed points of $\mathcal{F}$ by successive iterations. Let z denote the iteration number.

Algorithm 1. ***(Successive iterations on $\mathcal{F}$):*** *1) Start with an initial probability distribution g^0. 2) Given g^{z-1} find $g^z = \mathcal{F}(g^{z-1})$ for $z = 1, 2, \ldots$. 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. Finally, at the end of this section we propose a practical approximation to Algorithm 1 that is likely to act as a regularization.

Since one cannot store general continuous densities on a computer, only *approximate* iterations on $\mathcal{F}$ are feasible. In the next subsection we discuss discrete distributions. Then we discuss step function approximations of continuous densities. Finally we discuss approximations of continuous densities using a given parametric family.

3.1 The discrete case

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 a matrix Π with a typical element $\pi_{ij} = p_{Y|\theta}(\overline{Y}_j; \overline{\theta}_i)$, the vector p_Y in this section has in the j-th element $p_Y(\overline{Y}_j)$. We write $g(\overline{\theta}_i) = g_i$. In the discrete case equation (1) specializes to

$$\Pi' g_\theta = p_Y \tag{3}$$

for some discrete distribution g_θ. The definition of the mapping $\mathcal{F}$ (2) specializes to

$$\mathcal{F}(g)_i \equiv \sum_j \frac{\pi_{ij} g_i}{\sum_k \pi_{kj} g_k} \, p_Y(\overline{Y}_j) \quad \text{for all } i = 1, ..., N. \tag{4}$$

As we stated before, *existence* of a distribution g_θ that solves (3) is an issue. Since we assume throughout that Π is invertible and since $\sum_{i=1}^N g_{\theta,i} = 1$ is guaranteed,[7] all we need to assume in addition, to ensure existence, is that the vector $g_\theta = (\Pi')^{-1} p_Y$ has only non-negative elements.

A trivial adaptation of Proposition 1 guarantees necessity, therefore if g_θ solves (3) then g_θ is a fixed point of $\mathcal{F}$. The following proposition guarantees sufficiency.

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 (3).*

Since invertibility of Π implies completeness this proposition follows from Proposition 3.[8]

The following proposition guarantees that the successive iterations algorithm is locally stable under some conditions:

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

[8] In the Online Appendix we also provide a direct proof based on linear algebra.

Proposition 5. ***(Convergence)*** *Assume that a solution to the inverse problem g_θ exists. Assume that i) Π is invertible, ii) $g_{\theta,i} > 0$ for all i, and iii) $p_Y(\overline{Y}_j) > 0$ for all j.*

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 iterations on $\mathcal{F}$ converge locally to g_θ. Formally, there is an open neighborhood $S \subset \Theta$ 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.

Assuming $g_{\theta,i} > 0$ for all i is a mild requirement. It is clear that the set of Π's and p_Y's that imply $g_{\theta,j} = 0$ for some j is of measure zero, since in the discrete case (3) implies $g_\theta = (\Pi')^{-1} p_Y$.

However, the requirement that a fixed point satisfies $g^* > 0$ in Proposition 4 is very important: 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 fixed point with $g_i^* = 1$ for any 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. Therefore one has to design algorithms that keep the iterations away from these 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.[9] The

[9] 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 "wrong" 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

conjugate approximate algorithm that we use in the empirical application ensures that g^* is everywhere positive by construction.

A quick look at (3) may suggest that solving inverse 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, this makes matrix inversion unfeasible. In contrast, the algorithm of successive iterations on $\mathcal{F}$ completely sidesteps any matrix inversion. This plus the use of a conjugate approximate algorithm in subsection 3.3 below enables us to solve very high-dimensional problems.

3.2 Approximation in the continuous case

When θ and Y can take a continuum of values one can approximate the density by a class of functions with finite elements. In this subsection we rely on step functions to approximate the continuous distributions involved. We find conditions guaranteeing that the fixed points of this modified problem converge to a solution of the continuous inverse equation (1) as the step size $\varepsilon \to 0$. 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 (1) arbitrarily well.

In the text we give some details about how to build the approximating step functions, we leave the full details for the Appendix. We denote as an "$\varepsilon-$partition" a partition of $\mathcal{Y} \subset \mathcal{R}^M$ into $N_\varepsilon < \infty$ non-overlapping intervals $\mathbf{Y}_j^\varepsilon$ (more specifically, multidimensional intervals) covering the whole space, in other words satisfying $\mathcal{Y} \subset \cup_{j=1}^{N_\varepsilon} \mathbf{Y}_j^\varepsilon$. All finite sides of the intervals $\mathbf{Y}_j^\varepsilon$ must have length less than $\varepsilon > 0$. We partition Θ into the same number N_ε of analogous non-overlapping intervals $\boldsymbol{\theta}_j^\varepsilon$,

shown that in the vicinity of such points Newton's algorithm moves particularly slowly. 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.

although we require these to be compact. We form a probability vector p_Y^ε with the N_ε elements, elements given by $p_{Y,j} = \int\limits_{\mathbf{Y}_j^\varepsilon} p_Y$, and we form an $N_\varepsilon \times N_\varepsilon$ matrix Π^ε with the typical element $\pi_{i,j}^\varepsilon$ obtained by integrating $p_{Y|\theta}$ over the intervals $\mathbf{Y}_j^\varepsilon \times \boldsymbol{\theta}_i^\varepsilon$. Clearly each row of Π^ε sums to 1.

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

$$\Pi^{\varepsilon\prime} g_\theta^\varepsilon = p_Y^\varepsilon \tag{5}$$

We assume for now that this solution 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 (1) 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 distribution 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 (5) satisfies $g_\theta^\varepsilon \geq 0$.

Then the limit of any convergent subsequence of $G_\theta^{\varepsilon_k}$ solves (1). 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 (1).

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.

Assuming uniqueness we have

Proposition 6. ***(Approximation by step functions)*** *If the (continuous) inverse equation (1) 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.

3.3 Approximate conjugate algorithm and accuracy check

Proposition 6 shows a precise sense in which convergence to continuous solutions can be obtained. Combined with Proposition 5 it also suggests an algorithm to find an approximate solution, namely, use successive iterations with $\mathcal{F}$ defined from $\Pi^\varepsilon, p_Y^\varepsilon$ for very small ε. 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. This is a well known problem in solving Fredholm equations.

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 a real life application 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 as long as the solution of the inverse equation (1) 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 $q : \Theta \rightarrow R^{\nu}$ be a function such that the moments $E_p(q(\theta))$ suffice to pin down any density $p \in \mathcal{G}$.[10]

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

1) Start with a $g^0 \in \mathcal{G}$

2) Given $g^{z-1} \in \mathcal{G}$ find $g^z \in \mathcal{G}$ that approximates $\mathcal{F}(g^{z-1})$. We do this in two steps.

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

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

Repeat 2) until convergence.

In words, we project each successive iteration on $\mathcal{F}$ back onto the family $\mathcal{G}$. To the extent that the true fixed point is not too far from this family, we can hope to get a reasonably good approximation.

The following result allows for huge gains in computational speed when $\mathcal{G}$ is conjugate. 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. [11] *Given any density g, for any function $q : \Theta \rightarrow R^{\nu}$ we have*

$$E_{\mathcal{F}(g)}(q(\theta)) = E_{p_Y}\left[E_{p^g(\cdot|Y)}(q(\theta))\right]. \tag{6}$$

This result suggests the following Monte Carlo procedure to compute the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$ required in Step 2.a above: i) Draw $\mathcal{M}$ realizations of Y from p_Y; ii) For each draw $\overline{Y}$ compute the posterior moments of θ using g^{z-1} as the prior, that is $E_{p^{g^{z-1}}(\cdot|\overline{Y})}(q(\theta))$; iii) approximate E_{p_Y} by averaging the posterior moments obtained in step ii) over the $\mathcal{M}$ draws. The key is that if $\mathcal{G}$ is a family of conjugate priors for $p_{Y|\theta}$ and if the moments computed in step ii) are available in closed form, then

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

[11] 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 (6) in the Appendix.

part 2.a) of the algorithm 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.

Accuracy

After performing the iterations we need to check the accuracy of the approximate solution $g^Z \in \mathcal{G}$ obtained in the last iteration. For our purpose it is not crucial to satisfy (1) exactly, since the prior densities p_Y a researcher may state for observables can only be indicative, so a reasonable approximation to p_Y should be acceptable.

We check accuracy by comparing a sample of draws from the left-hand side density of (1) with a sample of draws from p_Y. Draws from the left-hand side density are straightforward to obtain: 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 then compare moments or interval frequencies from arbitrarily large samples. We apply this procedure in our empirical application below: Figure 2 plots the quantiles of the prior on observables (shaded area) and the quantiles of the distribution of the observables implied by the approximate fixed point (continuous line), and in section 4.3 we compare these quantiles.

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 (1) 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 Application

We apply the above ideas to the VAR in Christiano et al. (1999) (CEE), designed to study the effects of monetary policy shocks on the U.S. economy. We first show how the results obtained with four standard priors used previously in the literature lead to very disparate results. If the authors had used another one out of these standard priors, they would have arrived at different conclusions about the effects of monetary policy. We argue that these standard priors do not reflect reasonable prior knowledge of the economy. Then we describe our prior about observables – the prior about initial growth rates – and we show the posterior obtained with our prior.

CEE estimate a VAR in levels with output (real GDP), prices, commodity prices, federal funds rate, total reserves, nonborrowed reserves and money, using quarterly data from 1965 to 1995.[12] Structural innovations are obtained with the Choleski decomposition for the above variable ordering. The monetary policy shock is the one corresponding to the federal funds rate.

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

$$y_t = \sum_{i=1}^{P} B_i \; y_{t-i} + \gamma + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma), \quad t = 1, ..., T. \tag{7}$$

The parameters of the VAR are $\theta = (B, \Sigma)$, where B is a matrix defined as $B = (B_1, ..., B_P, \gamma)'$. P is the number of lags. The initial values $y^o_{-P+1}, ..., y^o_0$ (the superscript o denotes 'observed data') are treated as fixed and the analysis conditions on them.

4.1 Results obtained with standard priors for VARs

We consider four standard alternative priors. First, we reproduce the results of CEE using their (implicit) noninformative prior $p(B, \Sigma) \propto |\Sigma|^{-\frac{N+1}{2}}$ (see e.g. Zellner, 1971, Ch.8), hence the posterior mean of B is the OLS estimate.

We add three standard informative priors for VARs that are commonly used. We refer to them as the "Minnesota" prior, the "Sims-Zha" prior and the "Dynare" prior.

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

All these three priors have Normal-Inverted Wishart form, i.e., they satisfy

$$p(\operatorname{vec} B|\Sigma) = \mathcal{N}(\operatorname{vec} M, Q \otimes \Sigma), \tag{8}$$

$$p(\Sigma) = \mathcal{IW}(S, v), \tag{9}$$

where $\mathcal{IW}$ denotes the Inverted Wishart density and M, Q, S, v are prior parameters of appropriate dimensions. All three priors use the same values of M, S, v and they differ only in the value of Q.

These priors are all centered at the Random Walk model for each variable, meaning that the matrix M in (8) has the value of 1 in the positions corresponding to the first own lag of each variable and 0 everywhere else. Such priors originate in Doan et al. (1984) and they are commonly used because they are known to greatly improve the forecasting power of a VAR.

We follow common rules of thumb when setting the remaining parameters. Namely, we set the parameters S, v in (9) using the "empirical Bayes" approach.[13] Then we build three versions of the parameter Q in (8). 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-Zha" 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).[14]

Figure 1 shows the responses of output to a monetary policy shock estimated

[13] 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 whole 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 have 116 observations in our sample, but, as is common, we set the degree of freedom parameter to a much lower value $v = 10$ in order to have a rather loose prior.

[14] 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 $\lambda_1 = 0.2$,

with these priors. Responses of the remaining variables are reported in the Online Appendix. To facilitate comparisons we display the posterior obtained with the noninformative prior of CEE as a shaded region in all plots.

Panels A to C illustrate that persistence differs dramatically depending on the prior on parameters used. The noninformative prior (in gray) produces a short-lived effect (the plotted 90% posterior probability range contains zero after about 10 quarters). The "Minnesota" prior in panel A produces similar persistence as the noninformative prior but narrower error bands. The "Sims-Zha" 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.

This shows that Bayesian VARs can produce very different results in this application.[15] Most researchers will find little reason to choose one or another alternative based on a priori grounds, because it is difficult to formulate and assess priors on VAR parameters directly. Furthermore, as we show in Figure 2, these priors on parameters imply priors about data behavior that no analyst would ever hold, hence it is not reasonable to advocate their use on the grounds that they may represent an analyst's belief. This is why we consider priors on observables instead.

4.2 Prior about initial growth rates

We now formulate our prior on observables p_Y: a prior about growth rates. We specify the prior on only few periods. To specify the prior on many periods ($t = 1, ...T$ or even $t = 1, ...\infty$) would completely determine (or even overdetermine) the value of

$\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 1$, $\mu_5 = 1$, $\mu_6 = 1$; and in the "Dynare" prior we take $\lambda_1 = 0.33$, $\lambda_2 = 1$, $\lambda_3 = 0.5$, $\lambda_4 = 10^5$, $\mu_5 = 2$, $\mu_6 = 5$.

[15] So do classical VARs - see footnote 1.

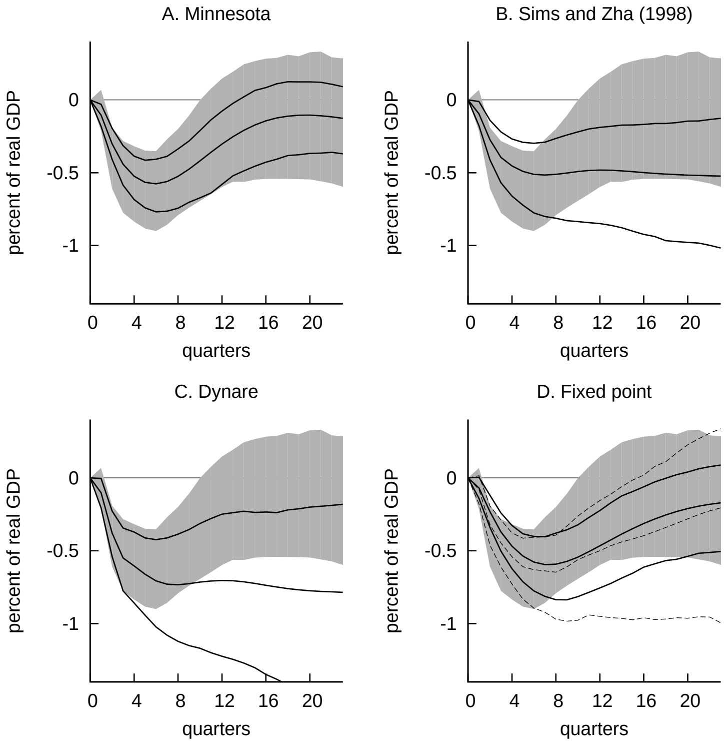

Figure 1 – Impulse response of output to a monetary shock: quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors. Gray area: quantiles 0.05 to 0.95 of the posterior obtained with the noninformative prior.

the coefficients so that the prior would completely dominate any sample information.

We specify our prior on the growth rates in the initial P periods conditional on the observed pre-sample values $y^o_{-P+1}, ..., y^o_0$. This prior is akin to the assumptions in the so-called "exact likelihood" approach (see Jarociński and Marcet (2010), Section 2 for a discussion) so it has the advantage of allowing to compare the results with this literature, which includes most frequentist small sample bias corrections.

Thus, we specify a $P \times N$ dimensional density $p_{\Delta y_1,...,\Delta y_P|y^o_{-P+1},...y^o_0}$ 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 $p_{y_1,...,y_P|y^o_{-P+1},...y^o_0}$. Ideally, the density we specify would be drawn from the purely subjective prior opinion of the user about the behavior of the variables. Future research can be directed at convenient ways of specifying such prior opinion. Instead, here we take an empirical Bayes approach and use the growth rates observed in the data to inform our prior.[16] 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 $p_{y_1,\dots,y_P|y^o_{-P+1},\dots y^o_0}$ 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 samples other than the estimation sample.

Figure 2 illustrates one aspect of a prior distribution of observables: the vertical axis shows the quantiles 0.05 and 0.95 of the densities of the observables y_t in periods $t = 1, 2, 3, 4$. The shaded regions show the quantiles of the prior density p_Y derived from the empirical prior discussed in the previous paragraph. For comparison, the dashed and dotted lines show the quantiles of the prior on observables implied by the standard informative priors for parameters of panels A, B and C in Figure 1.

These quantiles show that standard priors on parameters imply prior beliefs on observables that are unlikely to represent Bayesian prior information. For example, the Minnesota and the flat priors on parameters (used by CEE) are indistinguishable in this picture, they both show up as vertical lines: they both carry the “information” that output is very likely to grow by more than 100% in one period! This illustrates that, in contrast to our prior, standard informative priors for VARs often put much probability on unreasonable behavior of the observables in the first periods. Therefore, there is little reason to use these priors on parameters from a Bayesian point of view.

[16] See Morris (1983) for a classical reference on empirical Bayes or Efron (2010) for a more recent reference. 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. Our use of the empirical Bayes approach here follows Berger (1985, section 3.5.2) who suggests the data itself as a possible source of information about the marginal density of the data.

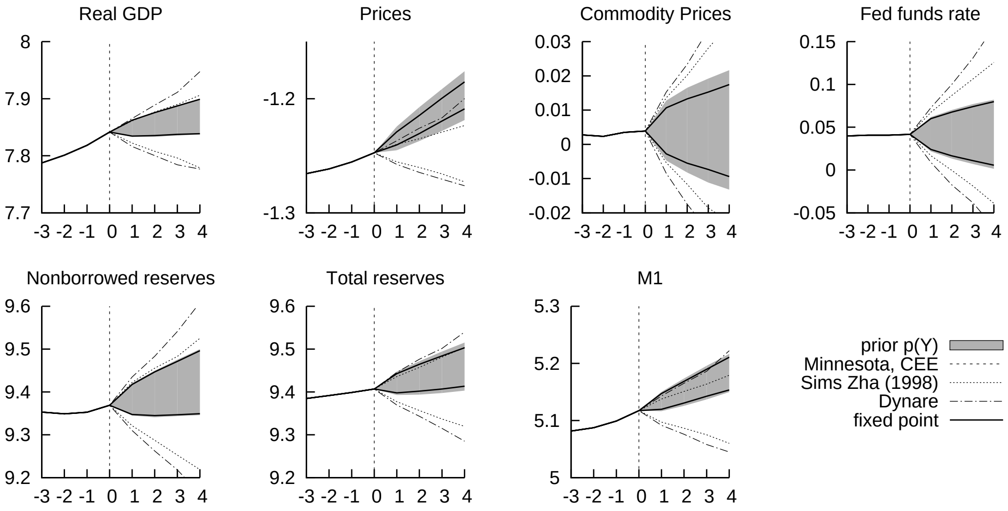

Figure 2 – Density of the observables implied by alternative priors. Quantiles 0.05 and 0.95 of the distribution in periods 1 to 4.

4.3 Results with the prior about initial growth rates

We implement the approximate conjugate algorithm proposed in section 3.3 when $\mathcal{G}$ is the family of Normal-Inverted Wishart densities (8)-(9). Using different random starting points g^0 for the algorithm[17] we find many different approximate fixed points with very similar implications for the observables. This happens because our prior about observables does not define a unique prior about parameters. Our prior states a distribution of dimension $NP = 28$, while the number of parameters for which it defines a prior p_θ is much larger.[18] 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 three informative

[17] To generate a starting point we draw one realization of the observables from the prior density of observables and compute the posterior density of the parameters based on this realization, using as a the prior the "Minnesota" prior from panel A *with randomly scaled parameters* Q, S, v. This posterior is used as the starting point g^0.

[18] 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.

priors used earlier.[19] We find 300 approximate fixed points that satisfy the restriction on $p(\Sigma)$. Finding each fixed point takes us from 2 to 5 minutes with Matlab on a standard personal computer.

From these 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, while maximum entropy can be interpreted as imposing as little prior knowledge as possible. It also happens to be the case that the maximum marginal likelihood fixed point has one of the lowest entropies, and that the maximum entropy prior 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 continuous lines in Figure 2 show the quantiles implied by the left hand side of (1) at a representative approximate fixed point with the restriction on $p(\Sigma)$. The continuous 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 and the restriction on $p(\Sigma)$, the approximate conjugate algorithm delivers a density of observables that is reasonable and close to the desired prior.[20]

The posterior for the fixed point with the highest marginal likelihood in the sample[21] is plotted with the continuous line in panel D of Figure 1. The posterior shows

[19] To find such priors we only iterate on M and Q, keeping the parameters S and v fixed and the same as in the standard informative priors for VARs used earlier.

[20] In the absence of the restriction on $p(\Sigma)$ we find fixed points for which the continuous 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 $p(\Sigma)$.

[21] This approach is frequently used in the applied literature to choose among competing priors. The marginal likelihood is $\int p(y^o|\cdot)p_\theta$, where y^o is the observed data. The log marginal likelihood of this prior is approximately 2780, compared with 2694, 2790 and 2783 respectively for the three

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%).[22] 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 than with a flat prior, implying that we have incorporated useful information in the estimation.

The dashed line in panel D of Figure 1 plots a posterior corresponding to the fixed point with the highest entropy.[23] It is comforting that this posterior confirms the main features of the highest marginal likelihood plotted with the continuous 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 main conclusion: that the response of output to a monetary policy shock is consistent with long-run neutrality of money and that the effect of a monetary shock is larger and more persistent than in CEE.

standard informative priors in panels A, B and C.

[22] 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.

[23] Entropy, defined as as $\int_\theta \log p(\theta) dp(\theta)$ measures the amount of information carried by a distribution. The log entropy of this fixed point equals -456, compared with -517, -779 and -664 respectively for the standard VAR priors in panels A, B and C. 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).

5 Conclusions

We have proposed using priors about observables in the estimation of a Bayesian VAR. Priors about observables are easier to interpret and, as shown by our empirical application to Christiano et al. (1999) they can make a difference in empirical work. 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 it converges. This algorithm works even in very high-dimensional problems that we consider.

In the empirical application we consider popular VAR priors give widely disparate results, sometimes imply non-neutrality of money in the long run. We show that these popular priors have odd implications for the prior on observables, they represent prior knowledge about observables that no reasonable economist would hold, hence they can not be justified from a Bayesian point of view. When we impose reasonable a priori behavior of observables, the posterior response of output to monetary policy shocks is larger and more persistent than under an uninformative prior and it is consistent with long-run neutrality of money.

Much future work can be based on the results here. 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. 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 9. More work is also needed on developing convenient approaches to formulate subjective priors on observables.

Appendix: Proofs

Proof of Proposition 1

We now show that when p_θ solves (1) then $\mathcal{F}(p_\theta) = p_\theta$. Clearly $\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g$ is non-zero whenever $p_Y(\overline{Y}) > 0$, 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 (1), 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

For any function $\widetilde{p}_\theta$ that satisfies (1) 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 from writing $p_{\theta|Y}$ in terms of Bayes' formula, the second because $\widetilde{p}_\theta$ satisfies (1).

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 $E(I_{Y^o}(Y) \mid \theta) = 0$. 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^*.

For a fixed point $g^* > 0$ 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 is in the Online Appendix

Proof of Proposition 5

The same reasoning as at the beginning of the proof of Proposition 3 guarantees that $\mathcal{F}(g_\theta)$ is well defined. By necessity and Proposition 4 g_θ is the unique fixed point of $\mathcal{F}$ where all elements are positive. Taking derivatives of $\mathcal{F}$ mechanically we have

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

Since g_θ solves the inverse equation and by assumption *iii)* we have $\sum_k \pi_{kj} g_{\theta,k} = p_Y(\overline{Y}_j) > 0$. Plugging this in the above expression and 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 have

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

so that

$$\frac{\partial \mathcal{F}(g_\theta)}{\partial g'} = I - \Delta^*. \tag{A.1}$$

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

$$\lambda_n \text{ is a real number and } 0 < \lambda_n \leq 1 \tag{A.2}$$

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

$$G^* \Delta^* = G^* \Pi \mathcal{D} \Pi' G^* \tag{A.3}$$

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.3) are invertible so that $G^*\Delta^*$ is invertible and none of its eigenvalues 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

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

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.2).

The eigenvalues of $(I - \Delta^*)$ are $1 - \lambda_n$, hence by (A.2) and (A.1) 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 iterations on $\mathcal{F}$ locally converge to g_θ. ■

Building ε-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 $N_\varepsilon < \infty$ (more specifically, multidimensional intervals) 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 $\text{supp}(\mathcal{Y})$ is 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 $\inf_{\text{supp}(\mathcal{Y}_m)} = Y_m^{\varepsilon,1} < Y_m^{\varepsilon,I_\varepsilon} = \sup_{\text{supp}(\mathcal{Y}_m)}$ where $\mathcal{Y}_m$ is the projection of the set $\mathcal{Y}$ 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 $i = 2, ..., I_\varepsilon - 1$ 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 $\inf_{\mathrm{supp}(\mathcal{Y}_m)} = -\infty$ (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 $\prod\limits_{m=1}^{M} (Y_m^{\varepsilon,i_m}, Y_m^{\varepsilon,i_m+1}]$ for some $i_m \in \{1, ..., I_\varepsilon - 1\}$, clearly $\mathcal{Y}$ is included in the union of these intervals. To construct sets such that $\mathbf{Y}_i^\varepsilon \subset \mathcal{Y}$ we overlap each interval with $\mathcal{Y}$, that is we set $\mathbf{Y}_i^\varepsilon = \mathrm{supp}(\mathcal{Y}) \cap \prod\limits_{m=1}^{M} (Y_m^{\varepsilon,i_m}, Y_m^{\varepsilon,i_m+1}]$ for all the intervals where the intersection is non-empty (empty sets have to be excluded to give a chance for Π^ε 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 for 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 $\theta_m^{\varepsilon,1} \to -\infty$ as $\varepsilon \to 0$. This guarantees that all $\boldsymbol{\theta}_i^\varepsilon$ are compact and $\cup_{i=1}^{N_\varepsilon} \boldsymbol{\theta}_i^\varepsilon \nearrow \mathrm{supp}(\Theta)$ as $\varepsilon \to 0$.

Let π_{ij}^ε be the integral of the likelihood over the corresponding sets in the partition:

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

and let 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 $\cup_{i=1}^{N_\varepsilon} \mathbf{Y}_i^\varepsilon = \mathrm{supp}(\mathcal{Y})$.

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.

Proof of Lemma 1

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

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

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

$$\int_{-\infty}^{\overline{Y}}\left[\int_{\Theta} p_{Y|\theta}(\widetilde{Y};\cdot)dG_{\theta}^{\varepsilon_k}\right] d\widetilde{Y} \to \int_{-\infty}^{\overline{Y}}\left[\int_{\Theta} p_{Y|\theta}(\widetilde{Y};\cdot)dG_{\theta}\right] d\widetilde{Y} \text{ as } k \to \infty. \tag{A.4}$$

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

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

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

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 (5).

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

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

The equality follows from the fact that the intervals $\mathbf{Y}_i^{\varepsilon_k}$ are disjoint and (5). One has to be careful arguing for the convergence part in (A.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 (A.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

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

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 $\left| \int\limits_{Y^{\varepsilon_k}(\overline{Y})}^{\overline{Y}} p_Y \right| \to 0$ because of continuity of p_Y. The convergence part in (A.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 (A.4) 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 (1) holds for the distribution functions implied by the densities p_θ and p_Y.■

Proof of Result 1

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

$$= \int_{\mathcal{Y}} \left(\int_{\Theta} q(\overline{\theta}) \; p_g(\overline{\theta}|\cdot) \; d\overline{\theta} \right) \; p_Y = E_{p_Y} \left(\; E_{p^g(\cdot|Y)}(q(\theta)) \; \right) \tag{A.6}$$

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 (6). ■

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.

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

Canay, I. A., Santos, A., and Shaikh, A. M. (2012). On the testability of identification in some nonparametric models with endogeneity. CeMMAP working papers CWP18/12, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

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. volume 6, part 2 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*, number 1A, chapter 2, pages 65–148. Amsterdam: North-Holland.

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.

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.

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

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. (1986). Forecasting with Bayesian vector autoregressions - five years of experience. *Journal of Business and Economic Statistics*, (4):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.

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.

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

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.

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