Priors about Observables in Vector Autoregressions*

Marek Jarociński
European Central Bank

Albert Marcet
Institut d'Analisi Economico, CEP and CEPR

June 25, 2012

Abstract

We formulate a prior about observables in a vector autoregression (VAR) and then solve the deconvolution problem for the implied prior about VAR parameters. Formulating a prior about observables is more intuitive than formulating a prior about VAR parameters directly, because VAR parameters are hard to interpret. Our numerical algorithm for approximating the implied prior about parameters works well even in high-dimensional problems and can be applied also for models other than VARs. In the empirical application we formulate a prior about growth rates of the observables in a VAR model of the United States economy. We find that this prior makes a big difference for the estimated persistence of output responses to monetary policy shocks, compared with the results of standard priors for VARs.

Keywords: Vector Autoregression, Bayesian Estimation, Prior about observables, Deconvolution, Monetary Policy Shocks

JEL codes: C11, C22, C32

*We thank Gianni Amisano, Manolo Arellano, Stéphane Bonhomme, Matteo Ciccarelli, Jean Pierre Florens, Oliver Linton, Bartosz Maćkowiak, Peter C.B. Phillips, Thomas J. Sargent, Frank Schorfheide and Chris Sims for useful conversations, and Harald Uhlig for his comments on the earlier version of this paper. All errors are our own. Albert Marcet acknowledges financial support from CREI, DGES (Ministerio de Educación y Ciencia), CIRIT (Generalitat de Catalunya) and European Community FP7-SSH grant MONFISPOL under grant agreement SSH-CT-2009-225149. The opinions expressed herein are those of the authors and do not necessarily represent those of the European Central Bank. Contacts: a.marcet@lse.ac.uk and marek.jarocinski@ecb.int.

1 Introduction

Vector autoregressions (VAR) are frequently estimated using the Bayesian approach. Formulating convincing priors for VARs is crucial, because in practice the sample is often small in relation to the number of estimated parameters and priors matter a lot for the results. However, formulating priors about VAR parameters is difficult because VAR parameters are hard to interpret.

We use a more intuitive approach to formulating the prior: we formulate the prior about observables. The contribution of this paper is twofold. First, we design a numerical algorithm that makes estimation with such priors feasible in VARs. Second, we present an empirical application where our approach allows us to interpret empirical evidence on the effect of monetary policy shocks.

The numerical algorithm that we introduce in this paper solves the following problem. When we formulate a prior about observables, a technical difficulty arises, because to use this prior in the estimation we need to first convert it into a prior about parameters. This involves solving a deconvolution problem, a Fredholm equation of the first kind. The existing methods for solving such problems are not applicable in large-scale problems, such as a typical VAR. Therefore, we design a new efficient numerical algorithm for solving this deconvolution problem. Our algorithm is applicable even when the prior about parameters has hundreds of parameters, like in VARs, and can be used also in models other than VARs.

In the empirical part of this paper we study the effects of monetary policy shocks in the U.S. with a VAR. We use a prior about growth rates of the variables in our VAR. With this prior we find that the effect of monetary policy on output is transitory, although more persistent than according to the Ordinary Least Squares (OLS) estimates. The fact that our prior is intuitive is crucial in this application. This is because when we use three most common versions of standard VAR priors (with default settings from the literature and econometric software) we find widely disparate results in this application. One of these priors produces similar results as ours, but the other two imply that the effect of monetary policy on output is permanent, and thus inconsistent with long-run neutrality of money present in most standard economic theories. Since these standard priors are statements about uninterpretable VAR coefficients, there is no way of assessing their plausibility. Thus, without a resort to our intuitive prior, it is difficult to make sense of these empirical results. There are two conclusions from this exercise. First, our results are consistent with long-run neutrality of money and discount the evidence against it. Second, this exercise shows that it is

useful and important in practice to formulate priors that can be interpreted and defended or criticized.

Related literature. In the existing Bayesian literature it is rare to formulate priors in terms of observables. An important early paper in this line of research is Kadane et al. (1980) who propose the 'predictive approach' to prior elicitation: formulating priors as predictive densities of observables. Kadane et al. (1996) apply this approach in a time series context. They use one period ahead forecasts and a univariate time series model. Their approach to solving the deconvolution problem is difficult to generalize to multiperiod distributions of observables in multivariate models, like VARs. A comprehensive theoretical discussion of using marginal densities of the observables to inform priors is in Berger (1985, Ch.3.5) but he discusses no algorithms for solving the deconvolution problem. The difficulty of solving the deconvolution problem in practice is a big obstacle to using priors about observables in applied work. The numerical algorithm that we propose in this paper is very general and can be used in all these cases in the literature.

Numerical approaches to solving inverse problems, including deconvolution problems, are surveyed in Carrasco et al. (2007). The techniques used in this literature and those used in calculus to solve Fredholm equations are usually designed to obtain very accurate solutions to relatively low-dimensional problems. They often involve solving non-linear systems of equations with gradient methods that would be unfeasible for our purpose since standard VARs involve very high-dimensional problems.

This paper is the first one, to our knowledge, that formulates and uses a prior about observables in a VAR. Some priors for VARs in the literature may appear to be priors about observables, while in fact they are priors about parameters. The popular Minnesota prior and dummy observation priors for VARs of Doan et al. (1984), Sims and Zha (1998) and others are examples of such priors. Another example is Villani (2009) who reparameterizes a VAR (assumed to be stationary) in terms of the steady state of the observables and formulates a prior about this steady state. Priors in all these approaches are stated about parameters and not as a prior density of observables, as in the present paper.

2 Finding the prior about parameters implied by a prior about observables

This section first states the implications of a prior about observables for the prior about parameters and then proposes a numerical algorithm for

approximating the implied prior about parameters.

2.1 The prior about observables and the implied deconvolution problem

Consider a model summarized in the density $p_{Y|\theta}$ that relates the distribution of the observable data Y to unknown parameters θ. To find the Bayesian posterior of θ we need to first formulate a prior p_θ. Formulating a subjective p_θ directly is difficult when θ lacks intuitive interpretation. Researchers often have much better intuition about the observable data Y than about the parameters θ. Therefore, in this paper we assume that the researcher formulates a prior density of the observables p_Y. The uncertainty represented in p_Y is a combination of the researcher's uncertainty about the actual values of parameters θ and the error terms of the model in $p_{Y|\theta}$.

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

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

This equation says that the joint density of observables Y and parameters θ, integrated over the parameters, has to equal the marginal density of Y as specified by the prior p_Y. Our task will be, given the known densities p_Y and $p_{Y|\theta}$, to find the prior density p_θ that satisfies the functional equation (1). Deconvolution problems of this type are known in calculus as Fredholm equations of the first kind and in statistics as inverse problems.

The above problem may not have any solution for some pairs of p_Y and $p_{Y|\theta}$. Obviously, some beliefs about the observables conveyed in p_Y can be incompatible with the model $p_{Y|\theta}$. However, as long as the density p_Y is at least approximately compatible with the model one will find a prior for parameters that approximately delivers the desired distribution p_Y. Therefore, nonexistence of an exact solution needs not be a problem in practice.

Another possibility is that the above problem has multiple solutions. This is likely to be the case e.g. when the dimension of θ is larger than the dimension of Y. In this case equation (1) delivers only a restriction on the prior. Then the researcher needs to complete the prior with a density of the parameters in the so far unrestricted directions. Therefore, multiplicity of solutions needs not be a problem in practice.

2.2 Fixed point formulation

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

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

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

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

$$p^g_{\theta|Y}(\overline{\theta}|\overline{Y}) \equiv \frac{p_{Y|\theta}(\overline{Y};\overline{\theta})\ g(\overline{\theta})}{\int_\Theta p_{Y|\theta}(\overline{Y};\cdot)\ g} \tag{3}$$

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

The following proposition gives the relationship between the fixed point of $\mathcal{F}_{p_Y}$ and problem (1):

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

Proof. We show if p_θ solves (1) then $\mathcal{F}_{p_Y}(p_\theta) = p_\theta$. We have for all $\overline{\theta} \in \Theta$

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

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 $p_{Y|\theta}$ is a density so it integrates to 1 over $\mathcal{Y}$. □

2.3 Approximate fixed point iteration with priors of a given functional form

Proposition 1 suggests a practical algorithm for solving the deconvolution (1): start with some density $g^0(\theta)$ and repeatedly apply mapping $\mathcal{F}_{p_Y}$, until the resulting density $g(\theta)$ stops changing. Proposition 1 states that this fixed point density is a good candidate solution of (1). However, iteration on the mapping $\mathcal{F}_{p_Y}$ is difficult because the densities g along the iteration do not in general have closed forms. Therefore, we implement an approximate iteration

in which at each step we restrict the density of the parameters $g(\theta)$ to be of a given convenient functional form. We place no restriction on the density p_Y except that it must be possible to generate draws of realizations from this distribution on a computer.

We first introduce some notation and definitions. Let $\mathcal{G}$ be a class of densities of θ of a given functional form. Let $h(\theta)$ be some function $h : \Theta \to R^m$, such that the moments $E_p(h(\theta))$ suffice to pin down the parameters of a density $p(\theta) \in \mathcal{G}$.[1]

Now we describe one approximate iteration. In this iteration, indexed by i, we take the density $g^{i-1} \in \mathcal{G}$ obtained in the preceding iteration and construct the density $g^i \in \mathcal{G}$ that approximates $\mathcal{F}_{p_Y}(g^{i-1})$. The iteration consists of two steps.

Step 1. We compute the moments $E_{\mathcal{F}_{p_Y}(g^{i-1})}(h(\theta))$ by Monte Carlo, based on the following result.

Result 1. *Given any g, for any function $h : \Theta \to R^m$ we have*

$$E_{\mathcal{F}_{p_Y}(g)}(h(\theta)) = E_{p_Y}\left[E_{p^g(\cdot|Y)}(h(\theta))\right] \tag{4}$$

Proof. [2]

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

$$= \int_{\mathcal{Y}} \left(\int_{\Theta} h(\overline{\theta})\ p_g(\overline{\theta}|\cdot)\ d\overline{\theta}\right)\ p_Y = E_{p_Y}\left(\ E_{p^g(\cdot|Y)}(h(\theta))\ \right) \tag{5}$$

where the first equality follows by definition of $\mathcal{F}_{p_Y}(g)$, the second by Fubini and the third by definition of E_{p_Y}. This proves (4). □

This result immediately suggests the following Monte Carlo procedure to compute the moments $E_{\mathcal{F}_{p_Y}(g^{i-1})}(h(\theta))$: a) Draw $\mathcal{M}$ realizations of Y from p_Y. b) For each draw $\overline{Y}$ compute the posterior moments of θ, $E_{p^{g^{i-1}}(\cdot|\overline{Y})}(h(\theta))$. Here the choice of $\mathcal{G}$ matters for the speed of the algorithm. When $\mathcal{G}$ is conjugate for $p_{Y|\theta}$, then these moments will typically be available in closed form. When $\mathcal{G}$ is not conjugate then a separate Monte Carlo procedure is needed for each draw $\overline{Y}$ in order to evaluate the moments. c) Approximate

[1] For example, if $\mathcal{G}$ is the class of gaussian densities then $h(\theta)$ consisting of the identity function $\theta \to \theta$ and the quadratic function $\theta \to \theta\theta'$ is a valid choice of h, because the first and the second moment suffice to pin down the parameters a gaussian density.

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

E_{p_Y} by averaging the posterior moments obtained in step b) over the $\mathcal{M}$ draws.

Step 2. We approximate the density $\mathcal{F}_{p_Y}(g^{i-1})$ by a density $g^i \in \mathcal{G}$ that has moments $E_{\mathcal{F}_{p_Y}(g^{i-1})}(h(\theta))$.

This completes one approximate iteration.

After performing the iterations we need to check how well g obtained with these iterations satisfies (1). We do it by comparing a sample of draws from the left-hand-side density of (1) with a sample of draws from the right-hand-side of (1). Generating a draw from the left-hand-side density is straightforward: we draw a realization of parameter values $\overline{\theta}$ from the density g, and then we draw Y from $p(Y|\overline{\theta})$. Generating a draw from the right-hand-side density is possible by assumption. The samples can be arbitrarily large and therefore the comparison of the underlying distributions can be arbitrarily precise.

2.4 Performance of the approximate iteration: a small Monte Carlo experiment

We do a small Monte Carlo experiment to check how the approximate iteration algorithm performs in a large-scale problem. We use the setup where problem (1) has a unique solution p_θ and we have a chance to uncover this solution with our algorithm, because we use the right functional form $\mathcal{G}$, i.e. such that $p_\theta \in \mathcal{G}$. We know the solution p_θ by the design of the experiment, but we proceed as if we only knew p_Y and $p_{Y|\theta}$, as in a real-life situation. We ask two questions. First, how difficult is it to find starting values for which the algorithm converges to the solution p_θ? Second, how precise and how fast is the algorithm?

The design of the Monte Carlo experiment is the following. We assume that the observables Y consist of 7 variables observed over 32 periods. We assume that these observables are generated by a $p_{Y|\theta}$ coming from a VAR model with 4 lags. We assume that the marginal density of the parameters p_θ is Normal-Inverted Wishart conjugate with the VAR model. The dimension of θ is 224 and a Normal-Inverted Wishart density of θ has 667 parameters. The marginal density of the data, p_Y is consistent with $p_{Y|\theta}$ and p_θ (according to (1)).

The results of the Monte Carlo experiment are promising. Using only the knowledge of p_Y we generate 100 random starting points for the algorithm and for each of these 100 starting values our algorithm converges. Each time it recovers the 667 parameters of p_θ with great precision in under 5 minutes. We provide all the details of this Monte Carlo experiment in Appendix B.

To our knowledge, our algorithm is not only robust and fast, but it is also the only feasible approach to finding these 667 parameters. For example, it would be impossible to find them by numerical optimization with gradient methods because the dimension of 667 is prohibitively large for such methods.

3 Empirical Application

In this section we study with a VAR the effects of monetary policy shocks on the U.S. economy. We first show the results obtained with standard priors for VARs used in the literature. Then we explain our prior formulated about observables, the prior about initial growth rates. Finally, we discuss the posterior results obtained with our prior.

The empirical exercise is the following. We reconsider the estimation of the effects of a monetary shocks in Christiano et al. (1999). They estimate a VAR with output, prices, commodity prices, federal funds rate, total reserves, nonborrowed reserves and money, using quarterly data from 1965 to 1995.[3] They orthogonalize the residuals with the Choleski decomposition of the variance of innovations with the above variable ordering. The monetary policy shock is the one corresponding to the federal funds rate. We follow Christiano et al. (1999) and deviate from them in only one aspect of the analysis, namely in that we use the Bayesian approach instead of their OLS estimation of the VAR.

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

$$y_t = \sum_{i=1}^{P} B_i \; y_{t-i} + \gamma + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma), \quad t = 1, ..., T. \tag{6}$$

$\mathcal{N}$ denotes the normal density. The parameters of the VAR are $\theta = (B, \Sigma)$, where B is a $K \times N$ matrix defined as $B = (B_1, ..., B_P, \gamma)'$, $K = NP + 1$, and Σ is an $N \times N$ symmetric positive definite matrix. As is common in the VAR literature we assume that the P initial values of the process are known and fixed at $(y^o_{-P+1}, ..., y^o_0)$, where superscript o denotes 'observed data'.

3.1 Results obtained with standard priors for VARs

We compute the posteriors with four alternative priors. First, we use the standard noninformative prior $p(B, \Sigma) \propto |\Sigma|^{-\frac{N+1}{2}}$ (see e.g. Zellner, 1971, Ch.8). With this prior, the posterior mean of the VAR parameters is equal

[3] See Christiano et al. (1999) for details. We downloaded these data from Christiano's webpage.

to the OLS estimate of Christiano et al. (1999). Next, we use three standard informative priors for VARs that are commonly used in applications. These three informative priors have Normal-Inverted Wishart form, i.e. they satisfy

$$p(\text{vec}\, B|\Sigma) = \mathcal{N}(\text{vec}\, M, Q \otimes \Sigma), \tag{7}$$

$$p(\Sigma) = \mathcal{IW}(S, v), \tag{8}$$

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

The standard informative priors for VARs are centered at the Random Walk model for each variable. In terms of (7) this means that the matrix M has the value of 1 in the positions corresponding to the first own lag of each variable and the value of 0 everywhere else (we denote this value of M by $\bar{M}$). Such priors are commonly used in VARs following Doan et al. (1984). The motivation is that the Random Walk model is both parsimonious (it corresponds to many VAR coefficients equal to zero) and persistent (it has a unit root). By pushing the posterior towards a parsimonious model we alleviate the well-known problem that VARs have too many parameters compared with the typical sample size in macroeconomics. By pushing the posterior towards a persistent model we alleviate the well-known problem that VARs estimated by OLS exhibit unreasonably little persistence (see e.g. Sims, 2000). Most researchers find this motivation convincing and there is little controversy around the choice of M.

The main challenge is to formulate the uncertainty about the Random Walk prior, i.e. the parameters Q, S, v. Formulating the uncertainty is challenging because VAR parameters lack intuitive interpretation. For example, we cannot tell if the value $B(1, 2) = 0.2$ is reasonable or not. To help applied researchers, the standard informative priors for VARs offer rules of thumb for setting parameters Q, S, v. We follow these rules of thumb. Therefore, first we set the parameters S, v in (8) to values $\bar{S}, \bar{v}$ obtained in an 'empirical Bayes' fashion.[4] Then we set the parameter Q by building it from a few scalar hyperparameters. Although it is easier to specify a few scalar hyperparameters than to specify a potentially large matrix Q directly, ultimately there is little intuitive guidance on the choice of these hyperparameters. Therefore,

[4]We use our dataset to estimate univariate autoregressive models with P lags for each of the variables in the VAR. Then we set S and v such that $E(\Sigma)$ is a diagonal matrix with error variances of the univariate autoregressions on the diagonal. We set the degree of freedom parameter $v = 10$. Our posteriors of the error variances of the univariate autoregression have the degree of freedom of 116, but we follow the common practice and use a value of v much lower than 116 in order to produce a rather loose prior.

we scan the literature and the documentation of popular econometric software and find three alternative recommendations for hyperparameter values. These three alternative hyperparameter values produce three versions of the parameter Q, denoted $\bar{Q}_1, \bar{Q}_2$ and $\bar{Q}_3$. $\bar{Q}_1$ corresponds to a version of the Minnesota prior close to Litterman (1986) and to baseline recommendations in the RATS manual (Doan, 2000). $\bar{Q}_2$ uses the Minnesota prior together with the 'dummy observations prior' and takes hyperparameter values from Sims and Zha (1998), which is the main academic reference for these priors. $\bar{Q}_3$ uses the Minnesota prior together with the 'dummy observations prior' and takes hyperparameter values used e.g. in Sims (2002) and implemented as the default in the Dynare package (Adjemian et al., 2011).[5]

Panels A, B and C of Figure 1 show the responses of output to a monetary policy shock estimated with the above priors. For brevity, we focus on the response of output. Responses of the remaining variables are reported in Appendix A. We plot quantiles 0.05, 0.5 and 0.95 of the posterior distribution of the impulse response. To facilitate comparisons we display the posterior obtained with the noninformative prior as a shaded region in all plots.

Figure 1 illustrates the main point of this subsection: persistence of the effect of monetary policy on output differs dramatically depending on the prior used. The noninformative prior produces a short-lived effect that disappears within about 16 quarters. The 'Minnesota' prior $(\bar{M}, \bar{Q}_1, \bar{S}, \bar{v})$ in panel A produces similar persistence as the noninformative prior. The 'Sims and Zha (1998)' prior $(\bar{M}, \bar{Q}_2, \bar{S}, \bar{v})$ in panel B and the 'Dynare' prior $(\bar{M}, \bar{Q}_3, \bar{S}, \bar{v})$ in panel C 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.

In the face of the disparate results presented in Figure 1 it is crucial to understand which of these four priors is the most reasonable, and then attach most weight to its implications. However, assessing these priors about VAR parameters is difficult because VAR parameters lack intuitive interpretation.

3.2 Prior about initial growth rates

We now formulate our prior about the observables p_Y: a prior about initial growth rates. This prior combines two elements: it is also a Random Walk

[5] In terms of Sims and Zha (1998) notation, $\bar{Q}_1$ is constructed using $\lambda_1 = 0.2$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 10^5$, $\mu_5 = 0$, $\mu_6 = 0$; $\bar{Q}_2$ is constructed using $\lambda_1 = 0.2$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 1$, $\mu_5 = 1$, $\mu_6 = 1$; and $\bar{Q}_3$ is constructed using $\lambda_1 = 0.33$, $\lambda_2 = 1$, $\lambda_3 = 0.5$, $\lambda_4 = 10^5$, $\mu_5 = 2$, $\mu_6 = 5$.

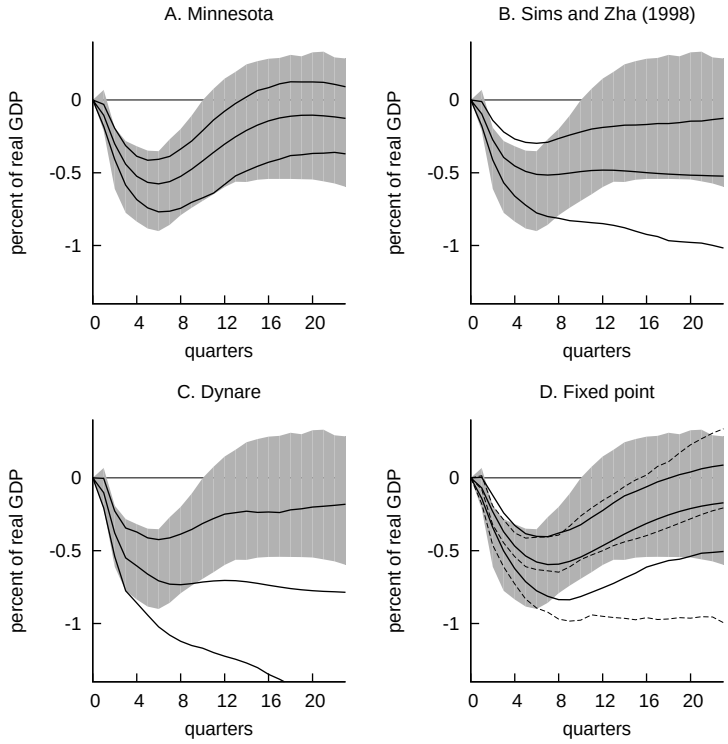

Figure 1 – Impulse response of output to a monetary shock: quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors. Gray area: quantiles 0.05 to 0.95 of the posterior obtained with the noninformative prior.

prior and it substitutes for the initial condition in an autoregressive model.

Our prior is a density of growth rates of the observables in periods 1 to P conditional on the observed pre-sample data

$$p(\Delta y_1, ..., \Delta y_P | y^o_{-P+1}, ...y^o_0) = f, \tag{9}$$

where f is a $P \times N$-dimensional density that we specify below. Note that the density of growth rates (9), coupled with the known value y^o_0 is, obviously, equivalent to a density of the levels of the observables

$$p(y_1, ..., y_P | y^o_{-P+1}, ...y^o_0) = f'. \tag{10}$$

First, our prior is a Random Walk prior. The Random Walk model implies that growth rates of the variables are independent and identically distributed in time and this is what our f conveys. We need to specify a prior distribution of the growth rate for each variable. This prior distribution can be subjective, but in this application we follow an Empirical Bayes approach

and use the sample to calibrate our prior. Namely, 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 our f the density of the observables implied by the posteriors of α_n, σ_n^2. This Empirical Bayes prior conveys the assumption that the first P observations behave, in terms of growth rates, similarly as the rest of the sample, which is a reasonable assumption in this application. Means and standard deviations of growth rates of the variables in the sample are reported in Table 1.

We are aware that the Empirical Bayes approach can be criticized for making the prior dependent on the data, but, in addition to its simplicity, we see three arguments for using it. First, the U.S. macroeconomic data studied here are very well known to researchers and it is virtually impossible to specify a truly data-independent subjective prior here, so we prefer to be explicit about the form of data-dependence. Second, the data inform our prior only through the very restricted auxiliary model. Third, the standard informative priors for VARs have Empirical Bayes elements anyway (in the specification of S), so our approach is in line with this literature.

Second, our prior substitutes for the initial condition in an autoregressive model. It is well known that the treatment of the initial values in autoregressive models is crucial in the Bayesian (as well as classical) criticism of the OLS estimator (see e.g. Sims, 2000). The common approach is to specify the likelihood function conditional on the initial P observations $y^o_{-P+1}, ...y^o_0$ and disregard the interdependence between these observations and model parameters. The alternative approach is to specify the 'exact' likelihood that includes also terms coming from the $P \times N$ dimensional density relating $y^o_{-P+1}, ...y^o_0$ and model parameters.[6] However, as is well known, specifying these terms is difficult when the model is potentially nonstationary. Our prior (9) is an elegant alternative way of specifying the $P \times N$ dimensional density relating $y^o_{-P+1}, ...y^o_0$ and model parameters.

We complement the above prior about observables with an additional restriction: that the marginal prior about Σ is the same as in the standard VAR priors, i.e. $\mathcal{IW}(\bar{S}, \bar{v})$. Without this restriction we find many different densities of (B, Σ) that are approximately consistent with (9), but put a lot of probability mass on small values of Σ, compensated by the large variance of B. We find these priors not to be reasonable and we believe more in the standard prior about Σ, $p(\Sigma) = \mathcal{IW}(\bar{S}, \bar{v})$.

Table 1 – Annualized growth rates of the variables: mean (standard deviation).

	1965-1995	1965-1985	1985-1995	1958-1964
Output	2.7	2.8	2.6	4.3
	(3.6)	(4.2)	(2.1)	(3.3)
Prices	5.0	5.9	3.1	1.8
	(2.5)	(2.4)	(1.1)	(1.3)
Commodity prices	0.0	0.0	0.2	0.0
	(2.1)	(2.2)	(1.8)	(0.7)
Fed funds rate	0.1	0.2	-0.2	0.2
	(4.8)	(5.6)	(2.1)	(1.3)
Non-borrowed reserves	5.3	4.3	7.4	1.5
	(9.1)	(8.8)	(9.3)	(5.8)
Total reserves	5.2	4.3	7.2	1.4
	(6.6)	(4.7)	(9.0)	(4.2)
Money (M1)	6.5	6.3	6.9	2.7
	(4.0)	(3.1)	(5.5)	(2.3)

3.3 Results with the prior about initial growth rates

We implement the algorithm from section 2.3 assuming that $\mathcal{F}$ is a class of Normal-Inverted Wishart densities (7)-(8). We keep the parameters S and v fixed at $\bar{S}$ and $\bar{v}$ respectively, and only iterate on M and Q. We generate 300 random starting points g^0 for the algorithm. To generate a starting point we draw one realization of the observables from (10) and compute the posterior density of the parameters based on this realization, using as a the prior a randomly scaled-up Random Walk prior. That is, the prior is Normal-Inverted Wishart with parameters $\bar{M}, \bar{Q}_1 \times 10^c, \bar{S}, \bar{v}$, where c is drawn from the uniform distribution on the interval $(0, 6)$. This posterior is used as the starting point g^0. About 90% of the starting points converge to fixed points that match quite well the desired prior about the observables.

Figure 2 reports quantiles 0.05 and 0.95 of various densities of the observables y_t in periods $t = 1, 2, 3, 4$. The continuous line shows the quantiles of the prior density (10). The dashed line shows the quantiles implied by one of the approximate fixed points. The dashed lines are quite similar to the continuous lines. This illustrates that the approximate fixed point approach delivers a density of observables that is reasonable and close to the prior in

[6]See e.g. Zellner (1971, Ch.7.1), Schotman and Van Dijk (1991a,b), Uhlig (1994) or Lubrano (1995).

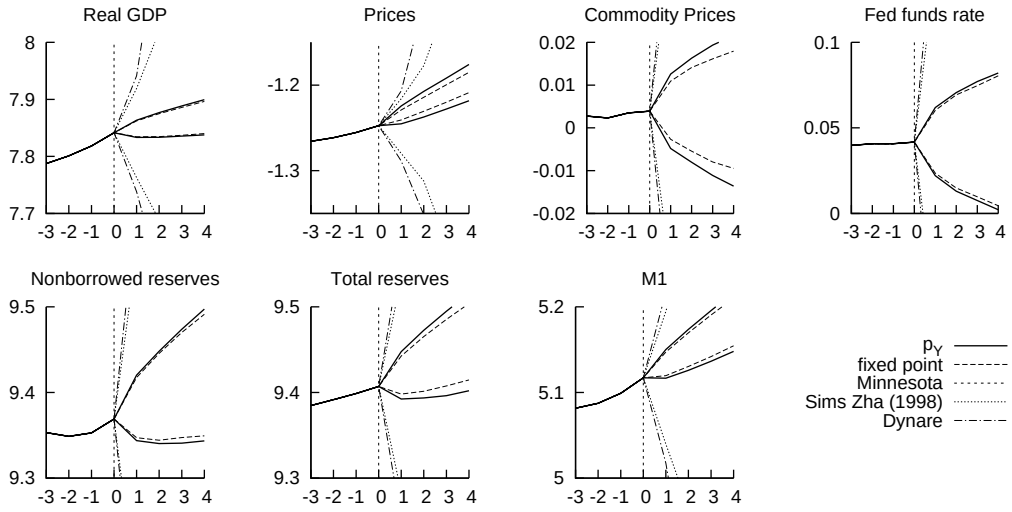

Figure 2 – Density of the observables implied by alternative priors. Quantiles 0.05 and 0.95 of the distribution in periods 1 to 4.

spite of the restrictions that we impose on the fixed point (namely, that it is a Normal-Inverted Wishart density with parameters $\bar{S}, \bar{v}$). The remaining lines show the quantiles implied by the standard informative priors for VARs. All these remaining lines quickly diverge towards very extreme values. This illustrates that the standard informative priors for VARs put much probability on unreasonable behavior of the observables in the first periods.

Our prior about observables (10) does not define a unique prior about parameters. Consequently, we find many different fixed-points that have basically the same implications for the observables in the first P periods. Therefore, to characterize their implications for the impulse response of output we plot two extreme cases in panel D of Figure 1. First, with the continuous line we plot the fixed point that produces the highest marginal likelihood in the analyzed sample.[7] Second, with the dashed line we plot the least informative fixed point - the fixed point with the highest entropy.[8] A researcher who wants to use the VAR model that delivers the best out-of-sample fore-

[7]The marginal likelihood is defined as $\int p(y|\theta)p(\theta)d\theta$. The log marginal likelihood of this prior is approximately 2780, compared with 2694, 2790 and 2783 respectively for the three standard informative priors in panels A, B and C.

[8]Entropy, defined as as $\int_\theta \log p(\theta)dp(\theta)$ measures the amount of information carried by a distribution. We derive the analytical expression for the entropy of the Normal-Inverted Wishart density using the results of Gupta and Srivastava (2010). 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.

casts would pick the fixed point with the highest marginal likelihood. A researcher who wants to impose as little prior knowledge as possible, beyond the restriction in (10), would pick the fixed point with the maximum entropy. Both of these impulse responses of output are more persistent than the responses obtained with the flat prior: output takes about 24 quarters to recover, instead of about 16 quarters with the flat prior. However, both of these impulse responses of output are mean-reverting, consistently with the long-run neutrality of money.

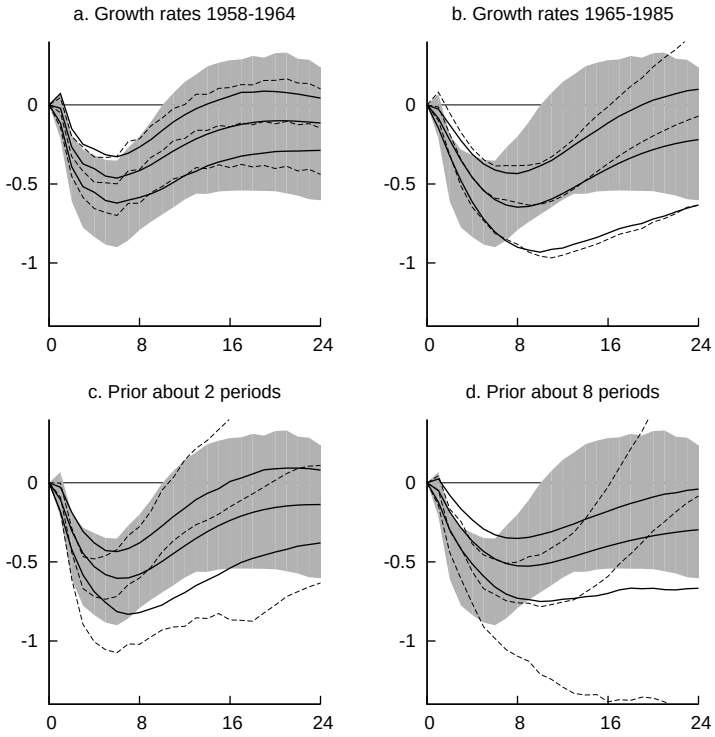

Figure 3 – Impulse response of output to a monetary shock: quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors about initial growth rates. Continuous lines: the fixed point with the highest marginal likelihood. Dashed lines: the fixed point with the highest entropy. Gray area: quantiles 0.05 to 0.95 of the posterior obtained with the noninformative prior.

Figure 3 reports the sensitivity of the posterior impulse responses of output to different specifications of the prior about the initial growth rates. In panel a. we calibrate the prior about growth rates, as well as the parameter $\bar{S}$, based on the data from the years 1958-1964, i.e. preceding the estimation sample 1965-1995. Thus, this prior uses no information from the estimation sample. As shown in panel a., when we use this prior, the response of output

is weaker and less persistent than in the baseline case (prior calibrated on the estimation sample).[9] In panel b. we calibrate the prior about growth rates based on the data from the years 1965-1985. In this case output response is more persistent than in the baseline case. In the next two experiments we deviate from the rule that our prior carries as much information as an initial condition in an autoregressive model. In panel c. we specify the prior about the first two growth rates only, Δy_1 and Δy_2. In this case output response is less persistent. In panel d. we specify the prior about the first 8 growth rates, Δy_1 up to Δy_8. In this case output response is more persistent but still it is not permanent, as with standard informative priors for VARs. The results hardly differ from the baseline (and we do not report them here for brevity) when we change the shape of f: when we use the empirical distribution of growth rates as our f (assuming that growth rates are independent across time and variables) or when we use as f a normal density. When we assume that the prior parameter v is 25 instead of 10, then the match of the density of the observables in the fixed point and the prior density of the observables f' becomes perfect, but the results (unreported here) hardly differ from the baseline.

Overall, we find that a range of reasonable priors about 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 not as persistent as some standard priors for VARs imply.

4 Conclusions

We have proposed a numerical algorithm for approximating the prior about parameters implied by a given prior about observables. This algorithm works even in high-dimensional problems like VARs. This algorithm makes it feasible and easy in practice to estimate VARs after formulating a prior about observables. This is useful, because priors about observables are easier to defend or criticise than priors about uninterpretable VAR parameters.

We have applied this approach to formulating priors in the study of the effects of monetary policy in the U.S. We have formulated a prior about initial growth rates of the observables in the VAR of Christiano et al. (1999).

[9] Growth rates in 1958-1964 were much less volatile than in the estimation sample 1965-1995 (see Table 1). Results are very similar to those in panel a. (we do not report them for brevity) also when we calibrate the prior using only post-1985 data. In the post-1985 period (called 'Great Moderation') output, prices and Fed funds rate are less volatile than in the main sample, while the monetary variables are more volatile than in the main sample (see Table 1).

Our posterior estimates are consistent with long-run neutrality of money and discount the evidence against neutrality implied by some popular VAR priors. This exercise shows the importance of using priors that can be interpreted and defended, or criticized.

Appendices

A Additional results for the monetary VAR

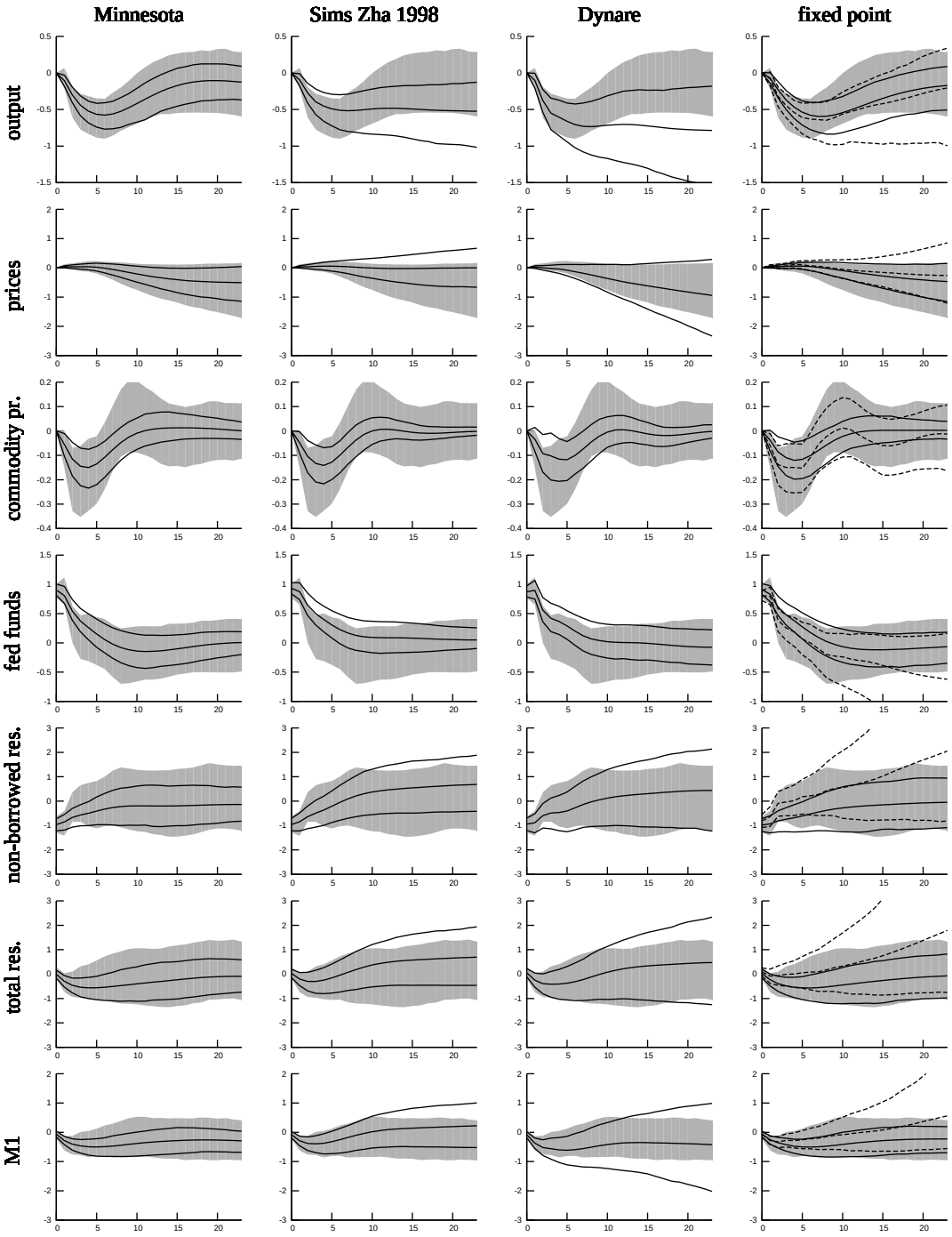

Figure 4 – Impulse responses of all variables to a monetary policy shock, quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors. Gray area: quantiles 0.05 to 0.95 of the posterior obtained with the noninformative prior.

B A Monte Carlo experiment with the approximate iterations

In this section we study by Monte Carlo the reliability of our algorithm for finding a prior about parameters given a prior about observables. We ask two questions of concern for a researcher who wants to implement our approximate iterations in practice: First, is it difficult to find starting values for which the algorithm converges to the solution of (1)? Second, how precise and how fast is the algorithm given a promising starting value? The results of the Monte Carlo experiment are promising. We generate 100 starting values, each obtained in a natural way from a random draw of Y from p_Y. We find that for each of these 100 starting values our algorithm recovers the 667 true parameters of p_θ^t with great precision in under 5 minutes.

B.1 The design of the experiment

The design of the experiment is based on the empirical application in section 3. We assume that the density of the data conditional on parameters $p_{Y|\theta}$ is given by the same VAR model with gaussian shocks,

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

We assume that the P initial values of the process are known and fixed at $(y_{-P+1}, ..., y_0)$, and starting from y_1 the process follows (B.1). The parameters of the VAR are $\theta = (B, \Sigma)$, where B is a $K \times N$ matrix defined as $B = (B_1, ..., B_P, \gamma)'$, $K = NP + 1$, and Σ is an $N \times N$ symmetric positive definite matrix. We assume that the 'true' marginal density of the parameters p_θ^t is Normal-Inverted Wishart, i.e. it satisfies

$$p(\text{vec}\, B|\Sigma) = \mathcal{N}(\text{vec}\, M, Q \otimes \Sigma), \tag{B.2}$$

$$p(\Sigma) = \mathcal{IW}(S, v), \tag{B.3}$$

where $\mathcal{N}$ denotes the normal density, $\mathcal{IW}$ denotes the Inverted Wishart density and M, Q, S, v are prior parameters of appropriate dimensions.[10] The density of (B, Σ) given in (B.2)-(B.3), model (B.1) and the initial value $(y_{-P+1}, ...y_{-1}, y_0)$ together determine p_Y - the density of y_t in $t = 1...T$. We would like to use values of (M, Q, S, v) and $(y_{-P+1}, ..., y_0)$ that are 'reasonable' and representative for potential real-life situations. Therefore, in

[10] We parameterize the Inverted Wishart density so that $E(\Sigma) = S/(v - N - 1)$.

this experiment we use the values of $(y^o_{-P+1}, ..., y^o_0)$ taken from the dataset of Christiano et al. (1999) (superscript o indicates 'observed data') and the values of M, Q, S, v that we found estimating model (B.1) on this dataset using the standard noninformative prior $p(B, \Sigma) = |\Sigma|^{-(N+1)/2}$.[11] There are $N = 7$ variables and $P = 4$ lags in this VAR. We set T, the number of periods in $p(Y)$, to 32. Note that if the dimension of the density $p(Y)$ is smaller than the dimension of $p(\theta)$ that we want to uncover, the uniqueness of the solution for $p(\theta)$ cannot be guaranteed. Setting $T = 32$ equalizes the two dimensions, since the dimension of Y is $TN = 224$, and the dimension of (B, Σ) (without counting the repeated entries in the symmetric matrix Σ) is also $KN + N(N+1)/2 = 224$.

B.2 Implementation of the approximate iterations

We set $\mathcal{G}$ to be the class of Normal-Inverted Wishart densities which are conjugate for the model (B.1), i.e. such that the posterior $p^g_{\theta|Y}$ is also Normal-Inverted Wishart. Our $h(\theta)$ consists of the identity function $\theta \to \theta$ and the quadratic function $\theta \to \theta\theta'$. Therefore, in Step 1 $E_{p^g_{\theta|\overline{Y}}}(h(\theta))$ is a vector of first and second moments of a Normal-Inverted Wishart density, which are available analytically. We average these moments over the $\mathcal{M} = 1000$ draws to obtain the moments $E_{\mathcal{F}_{p_Y}(g)}(h(\theta))$. In Step 2 we match the moments $E_{\mathcal{F}_{p_Y}(g)}(h(\theta))$ as well as possible with a Normal-Inverted Wishart density. Of course, a Normal-Inverted Wishart density cannot have arbitrary first and second moments because of its intrinsic restrictions, such as the Kronecker structure of the variance of B, so in general we cannot match $E_{\mathcal{F}_{p_Y}(g)}(h(\theta))$ exactly. Therefore, we just pick a subset of the first and second moments that results in convenient computations. We experimented with fitting a Normal-Inverted Wishart density to different sets of moments and we found many sets of moments that lead to similarly good convergence of the iterations in our application.

We run the algorithm 100 times. At each run we construct a random g^0 with the following procedure. We draw from $p(Y)$ a realization $\overline{Y}$. Then we compute the posterior of the parameters B, Σ conditional on $\overline{Y}$. This posterior belongs to $\mathcal{G}$ and we use the parameters M, Q and S of this posterior as the starting point. When computing the posterior we cannot use the noninformative prior because with only 32 observations the posterior would

[11] Specifically, define Y^{CEE} to be the $T^{CEE} \times N$ matrix collecting the observations on y_t from period 1 to T^{CEE} and define X^{CEE} to be the $T^{CEE} \times K$ matrix with the corresponding regressors: the lagged values of y_t and a column of 1s reflecting the constant term. Then we set $M = (X^{CEE\prime}X^{CEE})^{-1}X^{CEE\prime}Y^{CEE}$, $Q = (X^{CEE\prime}X^{CEE})^{-1}$, $S = (Y^{CEE} - X^{CEE}M)'(Y^{CEE} - X^{CEE}M)$ and $v = T^{CEE} - K - N - 1$.

be improper. Therefore, we use the standard Minnesota prior of Doan et al. (1984) cast as a Normal-Inverted Wishart prior (as e.g. in Kadiyala and Karlsson (1997)) and, to make it less informative, we blow up its standard deviation by 10^c where c is a random draw from a uniform distribution on (0,3). To introduce additional variation in the starting points, we draw v randomly from a uniform distribution between 10 and 200 (the 'true' v equals 81).

B.3 Results on the convergence of the iterations

The algorithm converges towards p_θ^t from each of the 100 starting points. To illustrate this, Figure 5 plots the evolution of g^i along the iterations for each starting point g^0. The first four panels show respectively the first element of M, the log determinant of Q, the log determinant of S and v. The values of these (functions of) g^i parameters are plotted against i with continuous lines. The 'true' values of these (functions of) parameters of p_θ^t are indicated with dashed horizontal lines. We see that in all plots the 100 continuous lines concentrate in the vicinity of the dashed line as iterations progress. We conclude that it is easy, in this application, to find good starting points for the algorithm based on the knowledge of p_Y alone. We also experimented with other starting points. For example, the algorithm also converges to p_θ^t when we start at the standard Minnesota prior or when we set M to a matrix of zeros. However, the algorithm runs into numerical problems or appears to stabilize away from p_θ^t when we change our good starting points selectively in only some dimensions, e.g. set a very tight density for the constant term γ in the VAR, or scale Q and S in opposite directions by factors of more than 100. The precision of the algorithm is very good. We report the precision in terms of the observables Y and not in terms of the parameters θ, because discrepancies of parameters from the 'true' values are hard to interpret. To illustrate the precision, the last panel shows the evolution of the Kullback-Leibler divergence between $p(Y)$ and $\int_\Theta p(Y|\theta)\ g^i(\theta)d\theta$[12] estimated from a sample of 1000 draws from each density. This plot suggests that already after about 20 iterations the discrepancies of g_θ^i from p_θ^t are negligible according to our estimator of Kullback-Leibler divergence. But what does this mean in practice? To illustrate the match of the distributions of the observables implied by g_θ^i and p_θ^t, Figure 6 plots the quantiles 0.05 and 0.95 of y_t against t for the 32 periods for which we specified p_Y. The continuous line shows the percentiles

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

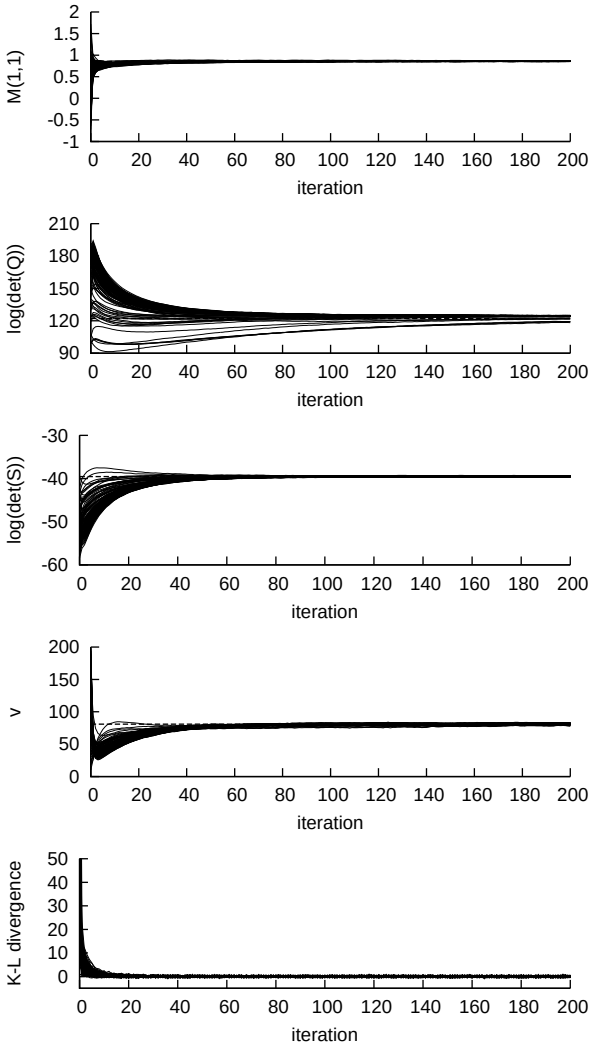

Figure 5 – Parameters of g^i along the iterations. Last plot: the estimated Kullback-Leibler divergence between $p(Y)$ and $\int_\Theta p(Y|\theta)\ g^i(\theta)d\theta$ along the iterations.

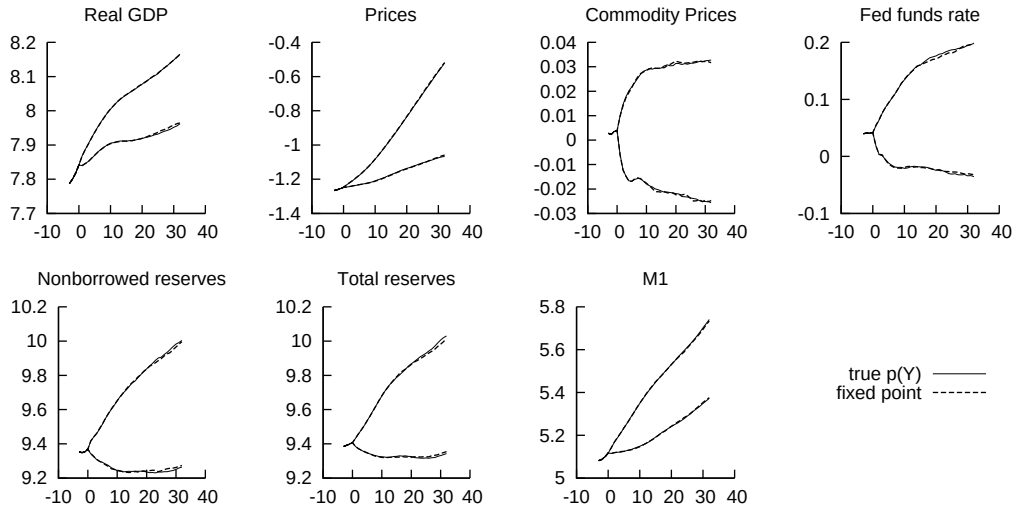

Figure 6 – Quantiles 0.05 and 0.95 of $p(Y)$ (continuous line) and $\int_\Theta p(Y|\theta)\ g^{200}(\theta)d\theta$ (dashed line) plotted against time.

of y_t generated from p_Y while the dashed lines show the percentiles of y_t generated from the distribution implied by g^{200}, $\int_\Theta p(Y|\theta)\ g^{200}(\theta)d\theta$ in the run of the algorithm witch achieved the largest Kullback-Leibler divergence from the target. We used 10,000 draws to reestimate the Kullback-Leibler divergences at the 200th iteration and to estimate the plotted quantiles. We see that the quantiles 0.05 and 0.95 of both distributions of Y basically coincide.

We conclude that the algorithm is extremely efficient compared to alternative approaches to such deconvolutions. In the current problem 200 iterations take under 5 minutes with matlab on a standard PC. Note that for a 7-variable VAR with 4 lags the dimension of M, Q, S, v (without counting the repeated entries in symmetric matrices) is $KN + K(K+1)/2 + N(N+1)/2 + 1 = 667$. To our knowledge, there are no other feasible approaches to finding these 667 parameters. For example, it would be impossible to numerically minimize an objective function (such as the Kullback-Leibler divergence between the left-hand-side and the right-hand-side of (1)) with gradient methods because the dimension of 667 is prohibitively large for such methods.

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.

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.

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.

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

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.

Kadiyala, K. R. and Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, 12(2):99–132.

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

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

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

Schotman, P. C. and Van Dijk, H. K. (1991a). A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics*, 49(1-2):195–238.

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

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

Sims, C. A. (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.

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

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

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.