Online Appendix to "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 Chair

February 15, 2013

A Additional results for the monetary VAR

This section reports additional empirical results that are mentioned in the main text. Figure A.1 reports responses of all variables to a monetary policy shock. Table A.1 reports growth rates of the variables in the main sample and in subsamples. Discussion of the sensitivity analysis follows.

*Contacts: albert.marcet@iae.csic.es and marek.jarocinski@ecb.int.

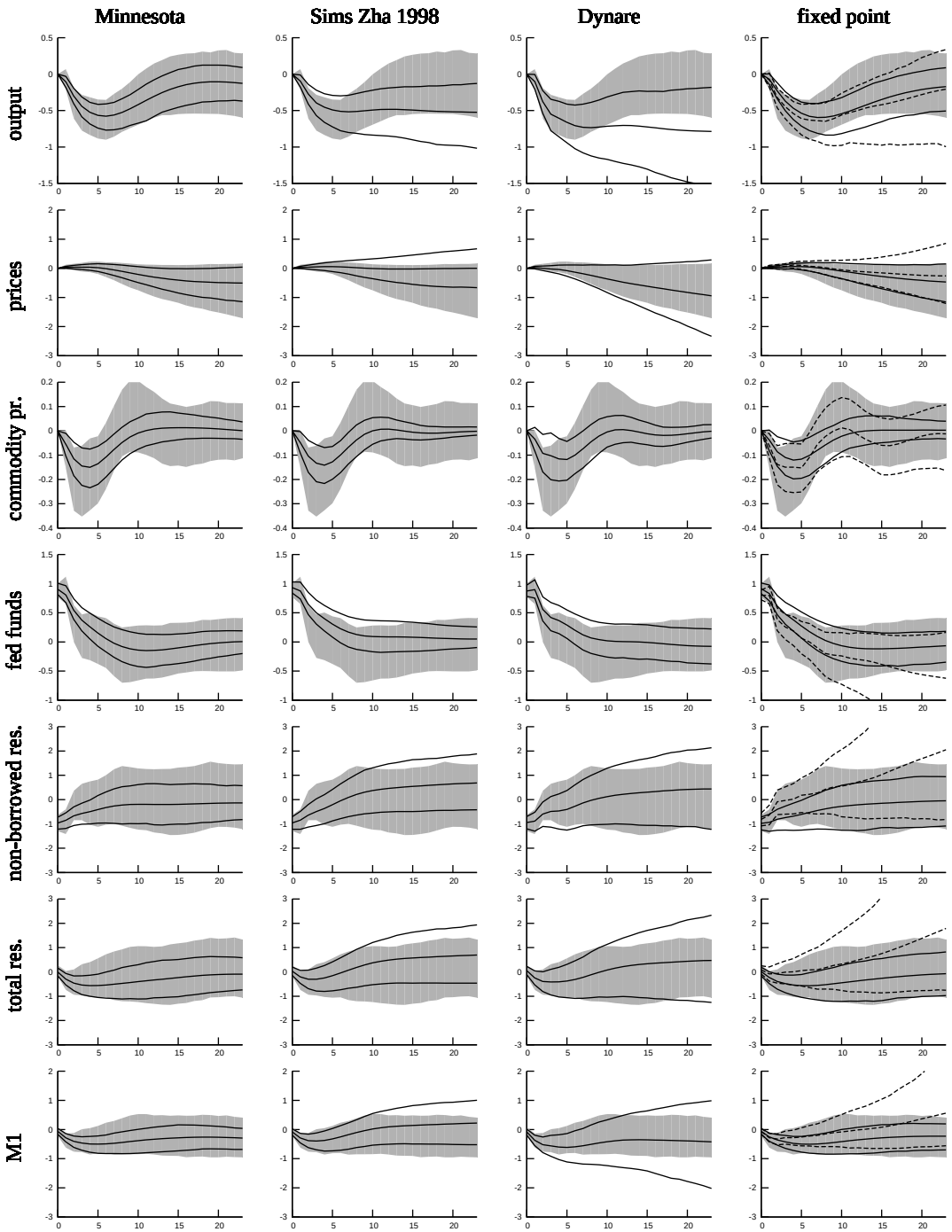

Figure A.1 – Impulse responses of all variables to a monetary policy shock, quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors. Gray area: quantiles 0.05 to 0.95 of the posterior obtained with the noninformative prior.

Table A.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)
Nonborrowed 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)

Figure A.2 reports the sensitivity of the posterior impulse responses of output to different specifications of the prior about the initial growth rates. When we discuss this figure below, our point of reference is the 'baseline' case, discussed in the paper, for which the prior about growth rates is calibrated on the estimation sample 1965-1995 and the posterior is reported in Figure 1.D of the paper.

In panel a. we calibrate the prior about growth rates, as well as the parameter S, based on the data from the years 1958-1964, i.e., preceding the estimation sample 1965-1995. As shown in panel a., when we use this prior, the response of output is weaker and less persistent than in the baseline. This prior uses no information from the estimation sample. This fact makes it more appealing on Bayesian grounds than

the baseline prior, which does use information from the estimation sample. However, this prior turns out to be very different from the baseline prior: it is very tight and centered around very different growth rates than those observed in the estimation sample. The reason is that growth rates in 1958-1964 (reported the last column of Table A.1) were quite different and much less volatile than in the estimation sample 1965-1995 (reported the first column of Table A.1). As shown in Table A.1, in 1958-1964 the standard deviation of the growth rate is 1.3 for prices (as opposed to 2.5 in the estimation sample), 0.7 for commodity prices (as opposed to 2.1), 1.3 for the fed funds rate (as opposed to 4.8), 5.8 for nonborrowed reserves (as opposed to 9.1), 4.2 for total reserves (as opposed to 6.6) and 2.3 for money (as opposed to 4.0). Only for output the difference is small (3.3 as opposed to 3.6). Some of the mean growth rates are also very different: 4.3 percent per annum for output (as opposed to 2.7), 1.8 for prices (as opposed to 5.0), 1.5 for nonborrowed reserves (as opposed to 5.3) etc. Results are very similar to those in panel a. (we do not report them for brevity) also when we calibrate the prior using only the so-called 'Great-Moderation' period, i.e., the post-1985 data. In the post-1985 data output, prices and Fed funds rate are also less volatile than in the main sample, while nonborrowed reserves, total reserves and money are more volatile than in the main sample (see the third column of Table A.1).

In panel b. we calibrate the prior about growth rates based on the part of the estimation sample before the 'Great Moderation', i.e., for the years 1965-1985. In this case output response is somewhat 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. Output response is less persistent than in the baseline. In panel d. we specify the prior about the first 8 growth rates, Δy_1 up to Δy_8. Now output response is more persistent than in the baseline.

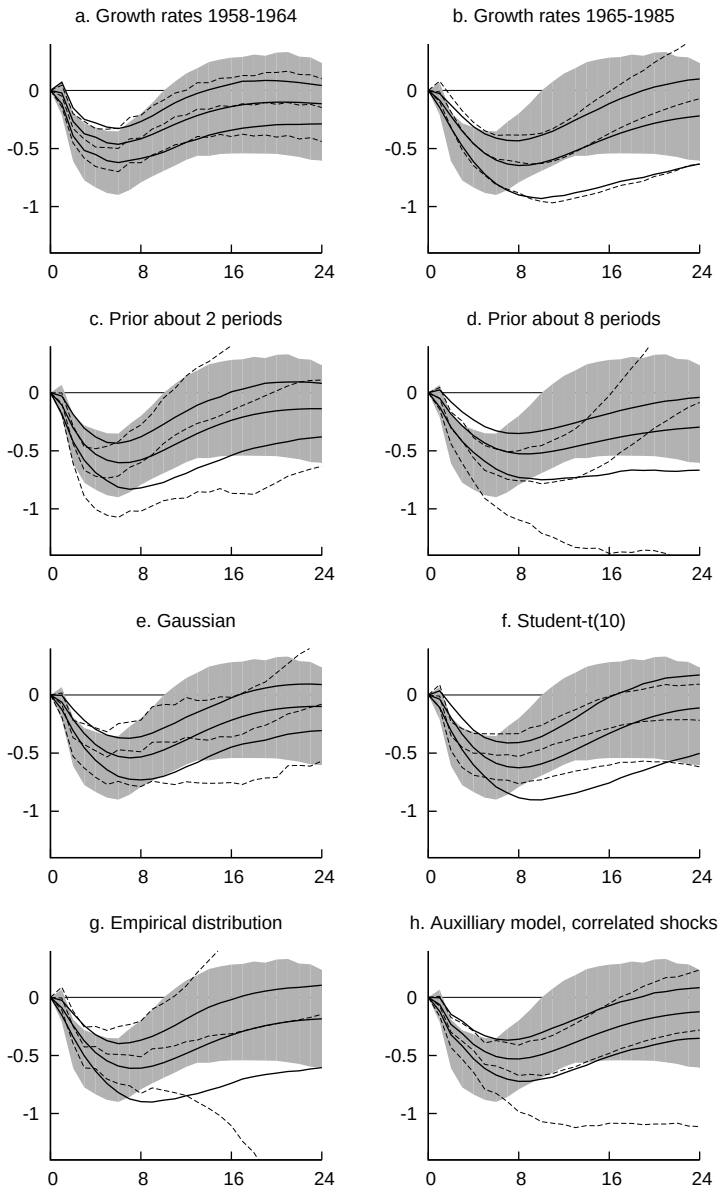

Figure A.2 – Impulse response of output to a monetary shock: quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors about initial growth rates. Continuous lines: the fixed point with the highest marginal likelihood. Dashed lines: the fixed point with the highest entropy. Gray area: quantiles 0.05 to 0.95 of the posterior obtained with the noninformative prior.

In panels e, f, g, h we keep the means and standard deviations of growth rates as in the baseline, while changing the shape of the prior. In panel e. the prior density of the observables is gaussian. Output responses are less persistent than in the baseline. In panel f. the prior density of the observables is Student-t with 10 degrees of freedom. Output responses are similar to the baseline. In panel g. we use as the prior the empirical distribution of growth rates in the sample (we simply draw observed growth rates with replacement). The maximum marginal likelihood responses are similar to the baseline, while the maximum entropy responses convey large uncertainty about medium and long run responses. Nevertheless, we do not rule out long-run neutrality of money. In panel h. we use the empirical Bayes prior with the the auxiliary model as in the baseline, except that shocks to growth rates are modeled as correlated across variables. Also in this case the maximum entropy responses convey much uncertainty about medium and long run, but do not rule out money neutrality.

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 but larger and more persistent than in CEE.

B A Monte Carlo experiment with the approximate conjugate algorithm

In this section we study by Monte Carlo the reliability of our approximate conjugate algorithm. The difference from the empirical application in section 4 of the main paper is that in the Monte Carlo we know the correct solution of the inverse problem (1) of the main paper. We ask two questions of concern for a researcher who wants to implement our algorithm in practice: First, is it difficult to find starting values for which the algorithm converges to the solution of the inverse problem (1) of the main

paper? Second, how precise and how fast is the algorithm? 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 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 4 of the main paper. We repeat all the assumptions here in order to make the present section self-contained.

We assume that the density of the data conditional on parameters $p_{Y|\theta}$ is given by the 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 $(y_{-P+1}, ..., y_0)$ are known 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 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.[1] The density of (B, Σ) given in (B.2)-(B.3), model (B.1) and the initial value of $(y_{-P+1}, ...y_{-1}, y_0)$ together determine p_Y – the density of y_t in $t = 1, ..., T$. We would like to use values of (M, Q, S, v)

[1] We parameterize the Inverted Wishart density so that $E(\Sigma) = S/(v - N - 1)$. See, e.g., Bauwens et al. (1999) Appendix A.2.6-A.2.7 for the properties of (B.2)-(B.3).

and $(y_{-P+1}, ..., y_0)$ that are 'reasonable' and representative for potential real-life situations. Therefore, in this experiment we use the values $(y^o_{-P+1}, ..., y^o_0)$ taken from the dataset of Christiano et al. (1999) (superscript o indicates 'observed data' as in Geweke (2005)) and the values of M, Q, S, v that we found estimating model (B.1) on this dataset using the standard noninformative prior $p(B, \Sigma) = |\Sigma|^{-(N+1)/2}$.[2] There are $N = 7$ variables and $P = 4$ lags in this VAR. We set T, the number of periods in $p(Y)$, to 33, because this choice of T equalizes the dimension of the density $p(Y)$ and the dimension of $p(\theta)$ that we want to uncover. The dimension of Y is $TN = 231$, and the dimension of (B, Σ) (without counting the repeated entries in the symmetric matrix Σ) is also $KN + N(N+1)/2 = 231$.

B.2 Implementation of the approximate conjugate algorithm

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 $p^g(\theta|\overline{Y})$ is also Normal-Inverted Wishart. Our $q(\theta)$ consists of the identity function and the quadratic function, $q(\theta) = (\text{vec}\,\theta', \text{vec}(\theta\theta')')$. We implement steps 2.a and 2.b of the algorithm in the following way.

In step 2.a given g^{z-1} we find the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$. We use the Monte Carlo outlined in the main paper. The Monte Carlo proceeds as follows.

i) We draw $\mathcal{M} = 1000$ realizations of Y from p_Y.

ii) Take one realization, $\overline{Y}$. We compute the posterior of B, Σ given data $\overline{Y}$ using g^{z-1} as the prior. Since g^{z-1} is conjugate Normal-Inverted Wishart, the posterior $p^{g^{z-1}}(B, \Sigma|\overline{Y})$ is also Normal-Inverted Wishart with the parameters given by standard formulas. Given this posterior, we compute and store its first and second moments,

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

$E_{p^{g^{z-1}}(\cdot|\overline{Y})}(q(\theta))$, which are also given by standard formulas. We repeat for each of the $\mathcal{M}$ realizations.

iii) We obtain the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$ using Result 1, i.e., as the average of the $\mathcal{M}$ moments computed in ii).

In step 2.b we match the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$ 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 $E_{\mathcal{F}(g^{z-1})}(q(\theta))$ exactly. Therefore, we just pick a subset of the first and second moments that we match exactly. We experimented with fitting a Normal-Inverted Wishart density to different sets of moments and we found that the choice of the set of moments is not critical: there are many sets of moments that lead to similarly good convergence of the iterations in our application. The Normal-Inverted Wishart density fitted to the moments $E_{\mathcal{F}(g^{z-1})}(q(\theta))$ is our next density, g^z.

We run the algorithm 100 times. At the beginning of 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 $\mathcal{G}$. When computing this posterior we cannot use the noninformative prior because with only 33 observations the posterior would be improper. Therefore, we use the "Minnesota" prior of section 4.1 of the main paper, but, 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 from each of the 100 starting points. To illustrate this, Figure B.1 plots the evolution of g^z 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^z parameters are plotted against z with continuous lines. The 'true' values of these (functions of) parameters of p_θ 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 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 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. In addition to the first four panels of Figure B.1 we also report the precision in terms of the observables Y, 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 $\int_\Theta p(Y|\theta)\ g^z(\theta)d\theta$ estimated from a sample of 1000 draws from each density.[3] This plot suggests that already after about 20 iterations the discrepancies of g^z_θ from p_θ are negligible as far as the implications for Y are concerned, 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^z_θ and p_θ, Figure B.2 plots the quantiles 0.05 and 0.95 of y_t against t for the 33 periods for which we specified p_Y. The continuous line shows the percentiles of y_t generated from p_Y while the dashed lines show the percentiles of y_t generated from

[3]We use $p(Y)$ as the weighting function in Kullback-Leibler divergence, i.e., we estimate $\int_\mathcal{Y} p(Y) \log\left(p(Y)\ /\ \int_\Theta p(Y|\theta)\ g^z(\theta)d\theta\right) dY$. We 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).

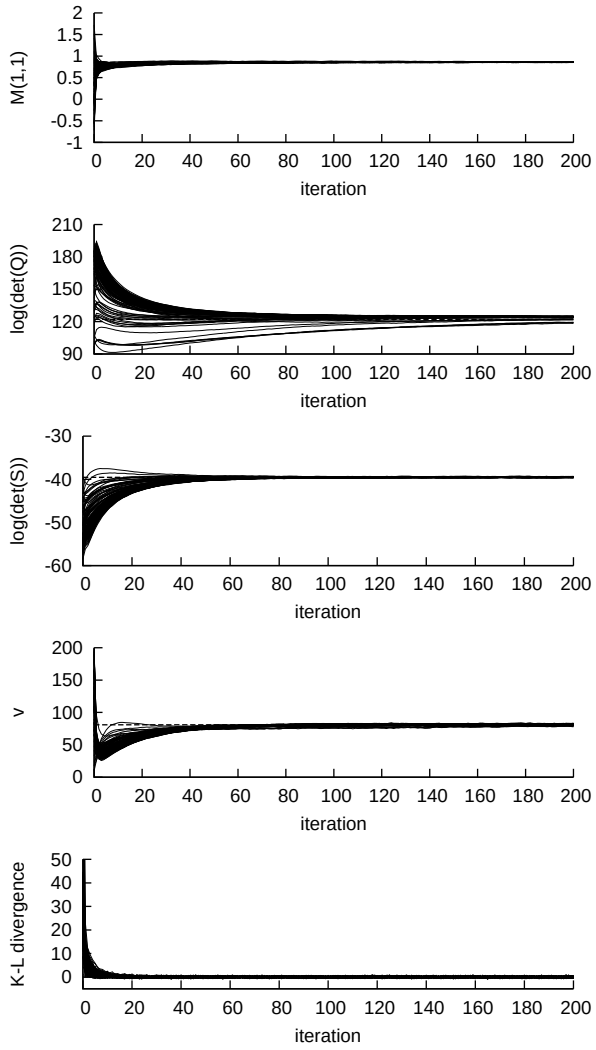

Figure B.1 – Parameters of g^z along the iterations. Last plot: the estimated Kullback-Leibler divergence between $p(Y)$ and $\int_\Theta p(Y|\theta)\ g^z(\theta)d\theta$ along the iterations.

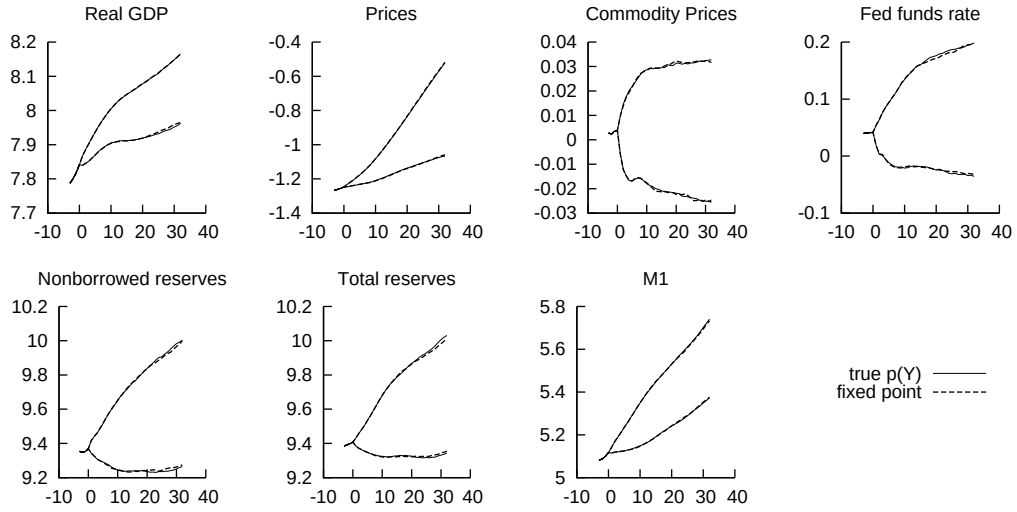

Figure B.2 – Quantiles 0.05 and 0.95 of $p(Y)$ (continuous line) and $\int_\Theta p(Y|\theta)\ g^{200}(\theta)d\theta$ (dashed line) plotted against time.

the distribution implied by g^{200}, $\int_\Theta p(Y|\theta)\ g^{200}(\theta)d\theta$, in the run of the algorithm that achieved the *largest* Kullback-Leibler divergence from the target, i.e., in the worst case. We used 10,000 draws of Y to reestimate the Kullback-Leibler divergences at the 200th iteration, in order to identify this worst case. We also used 10,000 draws of Y to estimate the plotted quantiles. We see in Figure B.2 that even in the case when the Kullback-Leibler divergence was the largest, 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 inverse problems. 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 the symmetric matrices Q and S) 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.

C Proof of a result from the main text

Proof of Proposition 4

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

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

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

Using $g_i^* > 0$, Lemma 2 and the fixed point condition imply that

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

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

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

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

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

so that g^* solves (3). ■

References

Bauwens, L., Lubrano, M., and Richard, J.-F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Advanced Texts in Econometrics. Oxford University Press, Oxford, first edition.

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.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. John Wiley and Sons, Hoboken, New Jersey.

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

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