Online Appendix to "Priors about Observables in Vector Autoregressions"

Marek Jarocinski
Albert Marcet

This version: June 2018
(February 2013)

Barcelona GSE Working Paper Series

Working Paper nº 685

Online Appendix to "Priors about Observables in Vector Autoregressions"*

Marek Jarociński
European Central Bank

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

June 24, 2018

Appendices A to C are in the main document. This document contains Appendices D to G.

D Implementation of the Approximate conjugate algorithm to a VAR

D.1 VAR and the Normal-Inverted Wishart prior

We restate some of the definitions already given in the main paper in order to make this appendix self-contained. The VAR model with Gaussian shocks is given by

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

The parameters of the VAR are $\theta = (B, \Sigma)$, where B is a $K \times N$ matrix defined as $B = (B_1, ..., B_P, c)'$, $K = NP + 1$, and Σ is an $N \times N$ symmetric positive definite matrix.

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

We will work with the conjugate Normal-Inverted Wishart priors for B, Σ given by

$$p(B, \Sigma) = \mathcal{NIW}(B, \Sigma;\ M, Q, S, v) \tag{D.2}$$

with parameters M, Q, S, v of dimensions $K \times N$, $K \times K$, $N \times N$ and 1×1 respectively. This density satisfies

$$p(\operatorname{vec} B|\Sigma) = \mathcal{N}(\operatorname{vec} M, \Sigma \otimes Q), \tag{D.3}$$

$$p(\Sigma) = \mathcal{IW}(S, v), \tag{D.4}$$

where $\mathcal{N}$ denotes the normal density, $\mathcal{IW}$ denotes the Inverted Wishart density parameterized 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 (D.2)-(D.4).

D.2 $\mathcal{G}$, moments of interest, initialization

We assume that the prior about observables is p_Y and let z denote the iteration number.

We set $\mathcal{G}$ to be the class of Normal-Inverted Wishart densities conjugate for the model (D.1), i.e., such that the posterior $p^g(\theta|\overline{Y})$ is also Normal-Inverted Wishart.

We specify the moments of interest $Eq(\theta)$. In the main paper we have specified the moments $M = E(B)$, $D = E(\Sigma^{-1})$, $\mathcal{V} = E(\operatorname{vec} B(\operatorname{vec} B)')$ and $\mathcal{H} = \operatorname{diag} E(\operatorname{vec} \Sigma^{-1} (\operatorname{vec} \Sigma^{-1})')$. Note that it is equivalent to work in terms of the two first moments plus the centered second moments, i.e. $M = E(B)$, $D = E(\Sigma^{-1})$, $V = \operatorname{cov}(\operatorname{vec} B)$ and $H = \operatorname{var}(\Sigma^{-1})$. In this appendix it is convenient to work in terms of the centered moments because some standard formulas are available for them. Througout this appendix, by 'var' we mean a matrix of variances and by 'cov' we mean the variance-covariance matrix. To be precise, for a matrix A, $\operatorname{var}(A)$ denotes the matrix in which each element is the variance of the corresponding element of A, $E\big((A - E(A))^2\big)$, while for a vector a, $\operatorname{cov}(a)$ is the variance-covariance matrix of a, $E\big((a - E(a))(a - E(a))'\big)$. In the above (and throughout this appendix) we use

the notation that for a matrix A, A^2 denotes a matrix in which each element is the square of the corresponding element of A.

We initialize the algorithm with a random g^0 obtained 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 use the Minnesota prior (described e.g. in the Appendix of the main paper), but, to make it less informative, we blow up its standard deviation by 10^u where u 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.

D.3 One iteration with the algorithm

This subsection explains in detail how we obtain $g^z(B, \Sigma)$ given $g^{z-1}(B, \Sigma)$. The assumptions are as follows.

1. After iteration $z-1$ the density of (B, Σ) is Normal-Inverted Wishart,

$$g^{z-1}(B, \Sigma) = \mathcal{NIW}\left(B, \Sigma;\ M^{z-1}, Q^{z-1}, S^{z-1}, v^{z-1}\right). \tag{D.5}$$

2. We have at our disposal a sample of J draws of Y from p_Y, denoted $\overline{Y}^j$, $j = 1...J$. Each drawn $\overline{Y}^j$ consists of values of N variables for T_1 periods.

Iteration z proceeds in three steps (we number the steps 2a, 2b and 3 consistently with the numeration in Algorithm 2 in the paper).

Step 2a. For each draw j we compute the posterior density of B, Σ conditional on $\overline{Y}^j$ with prior $g^{z-1}(B, \Sigma)$. This posterior density, denoted $p^z(B, \Sigma|\overline{Y}^j)$, is a Normal-Inverted Wishart density with parameters that we denote as $M^{z,j}, Q^{z,j}, S^{z,j}, v^{z,j}$ (with superscript z highlighting the dependence on the prior we are using in iteration z and superscript j highlighting the dependence on the drawn realization $\overline{Y}^j$). We also compute $R^{z,j} = (S^{z,j})^{-1}$. The form of the density and the closed-form expressions

for its parameters are given in Result D.1 below. The closed-form expressions for its moments of interest for us are given in the subsequent Result D.2. These results are known from the literature but we provide them here to keep this section self-contained.

Result D.1. *In the VAR given by (D.1) with the prior* $\mathcal{NIW}(B, \Sigma;\ M^{pri}, Q^{pri}, S^{pri}, v^{pri})$ *the posterior density is a Normal-Inverted Wishart density,*

$$p(B, \Sigma|Y) = \mathcal{NIW}\left(B, \Sigma;\ M^{po}, Q^{po}, S^{po}, v^{po}\right) \tag{D.6}$$

with parameters $M^{po}, Q^{po}, S^{po}, v^{po}$ *given by*

$$Q^{po} = \left(X^{po\prime}X^{po}\right)^{-1}, \quad M^{po} = \left(X^{po\prime}X^{po}\right)^{-1\prime} X^{po}Y^{po}$$

$$S^{po} = (Y^{po} - X^{po}M^{po})'(Y^{po} - X^{po}M^{po}) \text{ and } v^{po},$$

where

$$v^{po} = v^{pri} + T_1, \quad Y^{po} = \begin{pmatrix} Y^{pri} \\ Y \end{pmatrix}, \quad X^{po} = \begin{pmatrix} X^{pri} \\ X \end{pmatrix}.$$

Y^{pri} *and* X^{pri} *are defined as follows,*

$$Y^{pri} = \begin{pmatrix} (\text{chol}\, S^{pri})' \\ W^{pri}M^{pri} \end{pmatrix}, \quad X^{pri} = \begin{pmatrix} 0 \\ W^{pri} \end{pmatrix} \text{ and } W^{pri} = \text{chol}\left((Q^{pri})^{-1}\right)'. \tag{D.7}$$

Y *and* X *are defined as follows,*

$$\underset{T\times N}{Y} = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_T' \end{pmatrix} \quad \text{and} \quad \underset{T\times K}{X} = \begin{pmatrix} y_0' & y_{-1}' & \dots & y_{1-P}' & 1 \\ y_1' & y_0' & \dots & y_{2-P}' & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ y_{T-1}' & y_{T-2}' & \dots & y_{T-P}' & 1 \end{pmatrix}, \tag{D.8}$$

where T *is the number of observations in* Y*. In words,* X *is the matrix of lagged values of* Y *and exogenous variables (here: ones, corresponding to the constant term). Note that when constructing* X *we use the observed pre-sample values* $y_{-P+1}, ..., y_0$*, which are treated as given throughout the paper.*

Proof. It is easy to show that a Normal-Inverted Wishart density with parameters $M^{pri}, Q^{pri}, S^{pri}, v^{pri}$ can be equivalently written in terms of parameters Y^{pri}, X^{pri}, v^{pri},

$$p(B,\Sigma) \propto |\Sigma|^{-(v^{pri}+K+N+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(Y^{pri}-X^{pri}B)'(Y^{pri}-X^{pri}B)\Sigma^{-1}\right) \quad \text{(D.9)}$$

where Y^{pri} and X^{pri} are given in (D.7). Furthermore, The likelihood of Y is

$$p(Y|B,\Sigma) \propto |\Sigma|^{-T/2} \exp\left(-\frac{1}{2}\operatorname{tr}(Y-XB)'(Y-XB)\Sigma^{-1}\right). \quad \text{(D.10)}$$

Combining the prior (D.9) and the likelihood (D.10) in the standard way we arrive at the posterior given in (D.6). □

Result D.2. *When the density of B, Σ satisfies $p(B,\Sigma) = \mathcal{NIW}(B,\Sigma;\ M,Q,S,v)$, then Σ^{-1} and B have the following moments.*

$$E(\Sigma^{-1}) = R\,v, \quad \text{(D.11)}$$

$$\operatorname{var}(\Sigma^{-1}) = v\left(R^2 + \operatorname{diag}(R)\operatorname{diag}(R)'\right), \quad \text{(D.12)}$$

$$E(B) = M, \quad \text{(D.13)}$$

$$\operatorname{cov}(\operatorname{vec} B) = \frac{1}{v-N-1} S \otimes Q, \quad \text{(D.14)}$$

where $R \equiv S^{-1}$, R^2 is a matrix in which each element is the square of the corresponding element of R and $\operatorname{diag} R$ *is a column vector containing the main diagonal of R.*

Proof. (D.11) and (D.12) follow from the fact that the implied density of Σ^{-1} is Wishart

$$p(\Sigma^{-1}) = \mathcal{W}(R,v),$$

where $\mathcal{W}$ denotes the Wishart density and $R = S^{-1}$, and from the properties of the Wishart density. See e.g. Zellner (1971), p.389. (D.13) and (D.14) follow from the fact that the implied marginal density of B is Matricvariate Student

$$p(B) = \mathcal{T}(M,(Q)^{-1},S,v), \quad \text{(D.15)}$$

where $\mathcal{T}$ denotes the Matricvariate Student density and from the properties of the Matricvariate Student density. See Bauwens et al. (1999), Appendix A.2.7. □

Using these results we compute J realizations of the four moments of interest,

$$E_{g^{z-1}}(\Sigma^{-1}|Y^j) = R^{z,j}v^{z,j} \equiv D^{z,j}, \tag{D.16}$$

$$\text{var}_{g^{z-1}}(\Sigma^{-1}|Y^j) = v^{z,j}\left((R^{z,j})^2 + \text{diag}\, R^{z,j}\, \text{diag}\, R^{z,j\prime}\right) \equiv H^{z,j}, \tag{D.17}$$

$$E_{g^{z-1}}(\text{vec}\, B|Y^j) = \text{vec}\, M^{z,j} \equiv m^{z,j}, \tag{D.18}$$

$$\text{cov}_{g^{z-1}}(\text{vec}\, B|Y^j) = \frac{1}{v^{z,j}-N-1}S^{z,j} \otimes Q^{z,j} \equiv V^{z,j}. \tag{D.19}$$

Above we have also defined short-hand notations for the moments.

Step 2b. We approximate by Monte Carlo the moments of interest of B, Σ implied by $\mathcal{F}(g^{z-1}(B,\Sigma))$, denoted $E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$, $\text{var}_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$, $E_{\mathcal{F}(g^{z-1})}(B)$ and $\text{cov}_{\mathcal{F}(g^{z-1})}(\text{vec}\, B)$.

Result 1 of the main paper justifies the following Monte Carlo approximations:

$$E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1}) = \frac{1}{J}\sum_{j=1}^{J} D^{z,j}, \tag{D.20}$$

$$\text{var}_{\mathcal{F}(g^{z-1})}(\Sigma^{-1}) = \frac{1}{J}\sum_{j=1}^{J} H^{z,j} + \frac{1}{J}\sum_{j=1}^{J}(D^{z,j})^2 - E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})^2, \tag{D.21}$$

$$E_{\mathcal{F}(g^{z-1})}(B) = \frac{1}{J}\sum_{j=1}^{J} M^{z,j} \tag{D.22}$$

$$\text{cov}_{\mathcal{F}(g^{z-1})}(\text{vec}\, B) = \frac{1}{J}\sum_{j=1}^{J} V^{z,j} + \frac{1}{J}\sum_{j=1}^{J} m^{z,j}(m^{z,j})' - \text{vec}\, E_{\mathcal{F}(g^{z-1})}(B)\, \text{vec}\, E_{\mathcal{F}(g^{z-1})}(B)'. \tag{D.23}$$

The last two terms in the expressions for $\text{var}_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$ and $\text{cov}_{\mathcal{F}(g^{z-1})}(\text{vec}\, B)$ appear because variance is a central moment. These two terms give the variance of the posterior means of, respectively, Σ^{-1} and B, across draws. Hence, (D.21) and (D.23) are cases of the familiar expression for the total variance, obtained as the average conditional variance plus the variance of the conditional means.

Step 3. We find parameters of $g^z(B,\Sigma)$, denoted M^z, Q^z, S^z, v^z so as to match the moments $E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$, $\mathrm{var}_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$, $E_{\mathcal{F}(g^{z-1})}(B)$ and $\mathrm{cov}_{\mathcal{F}(g^{z-1})}(\mathrm{vec}\, B)$ as closely as possible with a Normal-Inverted Wishart density.

Note that $g^z(B,\Sigma)$ is an approximation of $\mathcal{F}(g^{z-1}(B,\Sigma))$ and there are many ways in which we could construct an approximation. After some experimenting we found that the approach described here works well in practice and is computationally convenient, but other approaches should be possible too. We proceed as follows. First, we find v^z and S^z that approximately match $E(\Sigma^{-1})$ and $\mathrm{var}(\Sigma^{-1})$ implied by $\mathcal{F}(g^{z-1}(B,\Sigma))$. Second, given v^z and S^z, we find M^z and Q^z that approximately match $E(B)$ and $\mathrm{cov}(\mathrm{vec}\, B)$ implied by $\mathcal{F}(g^{z-1}(B,\Sigma))$.

Step 3.1. We construct v^z, S^z that approximately match the moments of Σ^{-1} given by (D.20) and (D.21). We match (D.20) exactly and we match (D.21) approximately. Note that the variances in (D.21) can only be matched approximately because the Wishart density imposes a particular relation between the expectation and the variances and this relation might not hold between $E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$ and $\mathrm{var}_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$.

We construct v^z as

$$v^z = \frac{1}{N^2}\sum_{r=1}^{N}\sum_{c=1}^{N}\left(\widetilde{H}(r,c)\right), \tag{D.24}$$

i.e. as the average value of the entries of matrix $\widetilde{H}$, where

$$\widetilde{H} = \left((E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1}))^2 + \mathrm{diag}\, E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})\, \mathrm{diag}\, E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})'\right) \oslash \mathrm{var}_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})$$

and $\oslash$ denotes element-by-element division of two equally-sized matrices. To justify (D.24) observe that when Σ^{-1} comes from a Wishart distribution $\mathcal{W}(R,v)$ then, by the properties of the Wishart distribution,

$$\left((E(\Sigma^{-1}))^2 + \mathrm{diag}\, E(\Sigma^{-1})\, \mathrm{diag}\, E(\Sigma^{-1})'\right) \oslash \mathrm{var}(\Sigma^{-1}) =$$

$$\left(v^2R^2 + v^2\, \mathrm{diag}\, R(\mathrm{diag}\, R)'\right) \oslash v(R^2 + \mathrm{diag}\, R(\mathrm{diag}\, R)') = vU_N,$$

where U_N is an $N \times N$ matrix with all entries equal to 1.

We construct R^z as

$$R^z = E_{\mathcal{F}(g^{z-1})}(\Sigma^{-1})/v^z. \tag{D.25}$$

The above equation follows from the fact that if Σ^{-1} comes from a Wishart distribution $\mathcal{W}(R, v)$ then $E(\Sigma^{-1}) = Rv$. Given R^z, we compute

$$S^z = (R^z)^{-1}. \tag{D.26}$$

Step 3.2. We construct M^z, Q^z that approximately match the moments of B given by (D.22) and (D.23). We match $E_{\mathcal{F}(g^{z-1})}(B)$ exactly and we match $\mathrm{cov}_{\mathcal{F}(g^{z-1})}(\mathrm{vec}\, B)$ approximately. Note that the variance can only be matched approximately because the Matricvariate Student density with parameter S^z imposes on the covariance of B the structure $S^z \otimes C$, where C is some positive definite matrix, and $\mathrm{cov}_{\mathcal{F}(g^{z-1})}(\mathrm{vec}\, B)$ is not guaranteed to have such a structure.

We construct M^z as

$$M^z = E_{\mathcal{F}(g^{z-1})}(B). \tag{D.27}$$

We construct Q^z as

$$Q^z = \left(\sum_{n=1}^{N} V_n s_n\right) / \sum_{n=1}^{N} (s_n)^2, \tag{D.28}$$

where s_n denotes the nth element on the diagonal of S^z and V_n is the nth diagonal $K \times K$ block of $\mathrm{cov}_{\mathcal{F}(g^{z-1})}(\mathrm{vec}\, B) \times (v^z - N - 1)$, or, in other words, V_n is the variance of b_n, the vector of the coefficients of the nth equation of the VAR, multiplied by $(v^z - N - 1)$,

$$V_n \equiv \mathrm{cov}_{\mathcal{F}(g^{z-1})}(b_n)(v^z - N - 1).$$

The justification of (D.28) is the following. Note that we want to find Q^z such that $\mathrm{cov}_{\mathcal{F}(g^{z-1})}(\mathrm{vec}\, B) \approx \frac{1}{v^z - N - 1} S^z \otimes Q^z$, where the right-hand side is the variance of $\mathrm{vec}\, B$ implied by the Normal-Inverted Wishart density with parameters Q^z, S^z, v^z. First, to get rid of v^z we multiply by $(v^z - N - 1)$ and define $V = \mathrm{cov}_{\mathcal{F}(g^{z-1})}(\mathrm{vec}\, B)(v^z - N - 1)$,

so that $V \approx S^z \otimes Q^z$. Next, we want our approximation to be the best for the diagonal $K \times K$ blocks of V. Therefore, we specify the objective function to be minimized

$$\sum_{n=1}^{N} (s_n Q^z - V_n)^2, \tag{D.29}$$

i.e. the sum squared deviations between the N diagonal $K \times K$ blocks of V, denoted V_n for $n = 1, ..., N$ and the N diagonal $K \times K$ blocks of $S^z \otimes Q^z$, given by $s_n Q^z$. Note that the objective function is matrix-valued as there is a separate but analogous minimization problem for each entry of Q^z. The first order conditions of the minimizations are

$$2 \sum_{n=1}^{N} (s_n Q^z - V_n) s_n = 0,$$

implying

$$\sum_{n=1}^{N} {s_n}^2 Q^z = \sum_{n=1}^{N} V_n s_n.$$

Hence, (D.28) minimizes the objective function (D.29).

Summarizing, the parameters of $g^z(B, \Sigma)$, v^z, S^z, M^z, Q^z are given respectively by (D.24), (D.26), (D.27) and (D.28).

E A Monte Carlo experiment with the approximate conjugate algorithm

In this section we study by Monte Carlo the reliability of our approximate conjugate algorithm. In the Monte Carlo we assume that the solution of the inverse problem (4) is a known Normal-Inverted Wishart density. 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 (4)? 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.

E.1 The design of the experiment

The design of the experiment is based on the empirical application in section 4.2 of the main paper. We focus on this application because the dimension of the VAR is highest there.

We assume that the 'true' marginal density of the parameters p_θ is Normal-Inverted Wishart, with parameters M^*, Q^*, S^*, v^*. The density of (B, Σ) implied by M^*, Q^*, S^*, v^*, model (D.1) and an 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^*) and $(y_{-P+1}, ..., y_0)$ that are 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 (D.1) on this dataset using the standard noninformative prior $p(B, \Sigma) = |\Sigma|^{-(N+1)/2}$. More 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 $v^* = T^o - K - N - 1$.

There are $N = 7$ variables and $P = 4$ lags in this VAR. We set T, the number of periods in $p(Y)$, to 33, consistently with our informal rule of thumb. Namely, 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 $KN + N(N + 1)/2 = 231$.

E.2 Results on the convergence of the iterations

The algorithm converges towards p_θ from each of the 100 starting points. To illustrate this, Figure E.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 solid 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 solid 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 c 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 E.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. 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). The 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

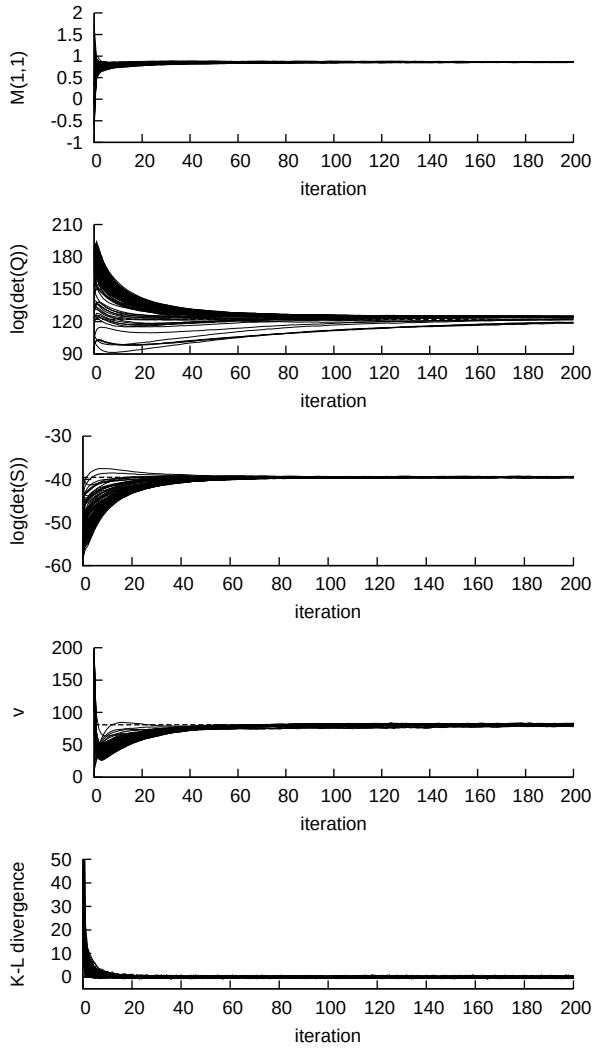

Figure E.1 – Parameters of g^z along the iterations. Last plot: the estimated Kullback-Leibler divergence between $p(Y)$ and $\int_\Theta p(Y|\theta)\ g^z(\theta)d\theta$ along the iterations.

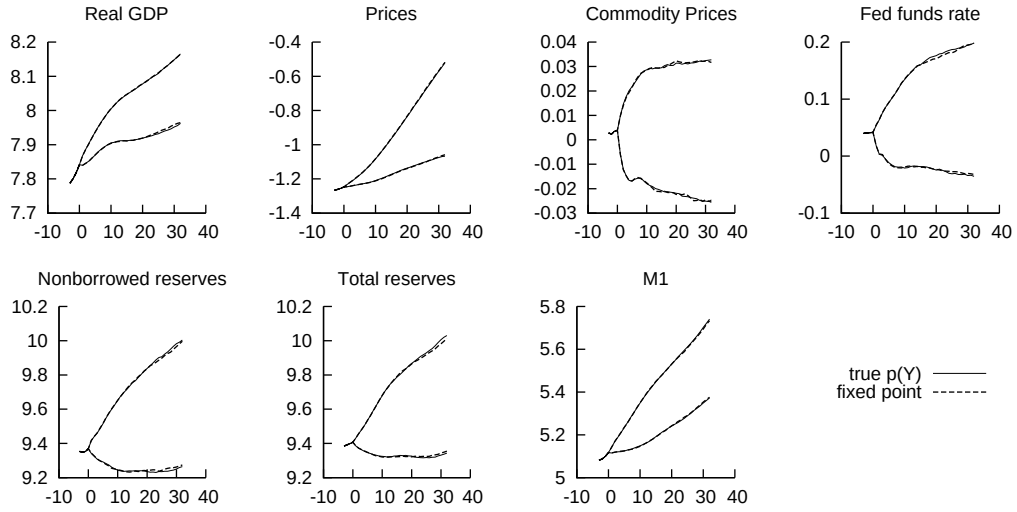

Figure E.2 – Quantiles 0.05 and 0.95 of $p(Y)$ (solid line) and $\int_\Theta p(Y|\theta)\ g^{200}(\theta)d\theta$ (dashed line) plotted against time.

distributions of the observables implied by g_θ^z and p_θ, Figure E.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 solid line shows the percentiles 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 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 E.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 (4)) with gradient methods because the dimension of 667 is prohibitively large for such methods.

F Additional results for the VAR of Christiano, Eichenbaum and Evans (1999)

This section reports additional empirical results related to the application in section 4.2. Figure F.1 reports responses of all variables to a monetary policy shock. Table F.1 reports growth rates of the variables in the main sample and in subsamples. Discussion of the sensitivity analysis follows.

Figure F.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 main paper, for which the prior about growth rates is calibrated on the estimation sample 1965-1995.

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 F.1) were quite different and much less volatile than in the estimation sample 1965-1995 (reported the first column of Table F.1). As shown in Table F.1, in 1958-

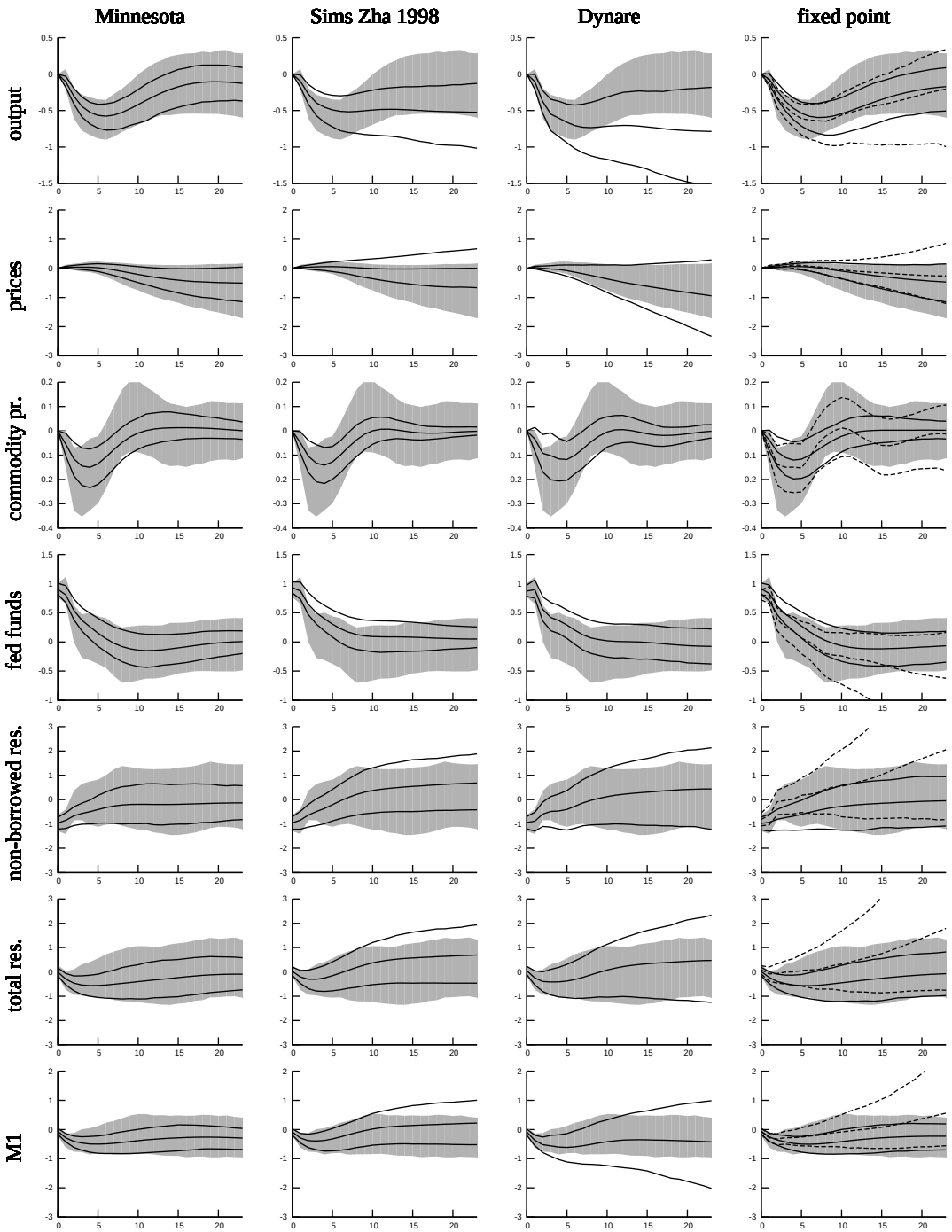

Figure F.1 – Impulse responses of all variables to a monetary policy shock, quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors. Shaded area: quantiles 0.05 to 0.95 of the posterior obtained with the noninformative prior.

Table F.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)

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 F.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.

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

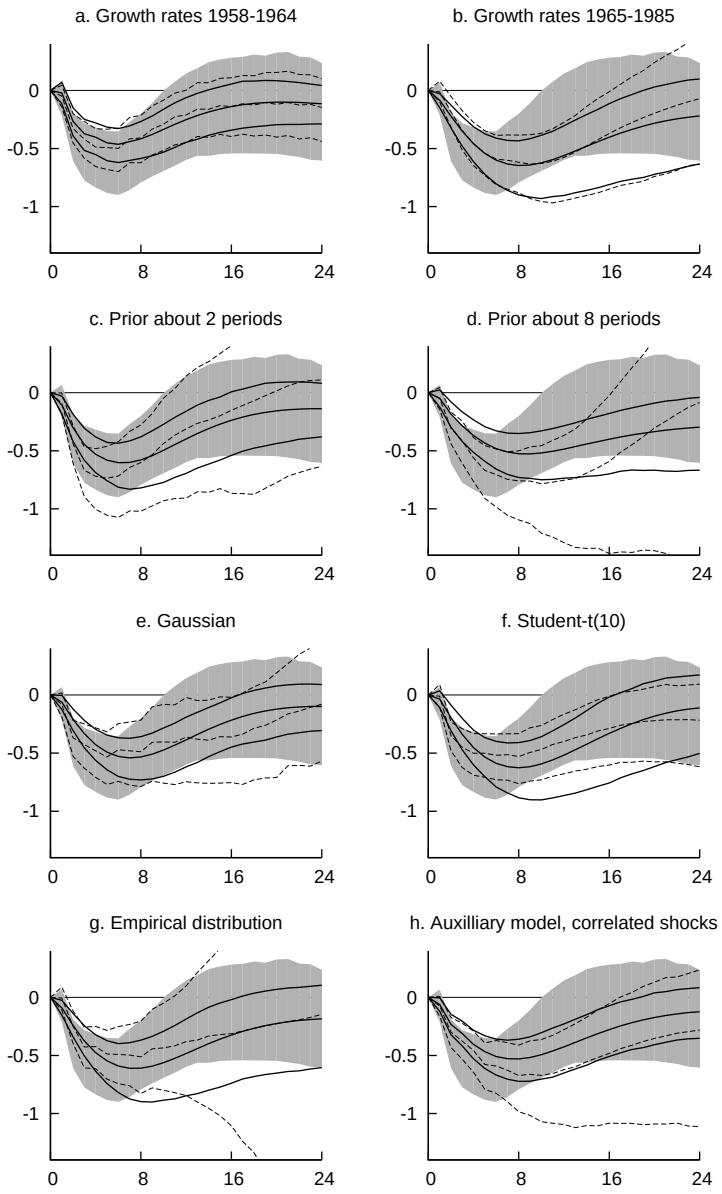

Figure F.2 – Impulse response of output to a monetary shock: quantiles 0.05, 0.5 and 0.95 of the posteriors obtained with alternative priors about initial growth rates. Solid lines: the fixed point with the highest marginal likelihood. Dashed lines: the fixed point with the highest entropy. Shaded area: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

consistent with long-run neutrality of money but larger and more persistent than in CEE.

G p_θ can not be found by a change of variable

This section discusses the approach to finding p_θ consisting of two steps: i) writing down the observables y as a function of model parameters θ and shocks, and ii) writing the density of observables p_Y and finding the implied prior p_θ using the change of variable technique. We argue that this approach is not promising.

To apply the change of variable formula we would need to go from the joint density of (Y, U) to the joint density of (θ, U). Note that there is no one-to-one mapping between observables Y and parameters θ, because the shocks U affect the observables too, so we need to include the densities of the shocks too. The problem is how to formulate the joint density of (Y, U). This joint density would need to be consistent with the prior about observables p_Y, with the assumed density of the shocks, and with the independence of parameters and shocks. Unfortunately, it is not clear how to formulate the joint density of (Y, U) satisfying these constraints and as a result this approach does not seem to be promising.

To illustrate these difficulties, consider the AR(1) model and suppose the researcher specifies a prior density of the observable in periods 1 and 2. In this case $\theta = \{\alpha, \rho\}$ and the mapping from (θ, U) to (Y, U) is as follows:

$$y_1 = \alpha + \rho y_0 + u_1 \tag{G.1}$$

$$y_2 = \alpha + \alpha\rho + \rho^2 y_0 + \rho u_1 + u_2 \tag{G.2}$$

$$u_1 = u_1 \tag{G.3}$$

$$u_2 = u_2 \tag{G.4}$$

It is easy to verify that the Jacobian matrix of this transformation is:

$$\begin{pmatrix} 1 & y_0 & 1 & 0 \\ 1+\rho & \alpha + 2\rho y_0 + u_1 & \rho & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The determinant of this matrix is $\alpha+(\rho-1)y_0+u_1$, and the absolute value of this term multiplies the distribution in the new parameter space (α, ρ, u_1, u_2). This term cannot be factorized into terms involving only us and terms involving only the parameters. Therefore, the obtained density will not, in general, be consistent with independence of the model parameters and errors.

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

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.

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