Online Appendix for: Bayesian Inference on Structural Impulse Response Functions

Mikkel Plagborg-Møller

August 26, 2016

This document contains supplemental material for the article "Bayesian Inference on Structural Impulse Response Functions". It contains a simulation study illustrating the workings of the SVMA procedure; supporting results and discussion for the news shock application; technical details concerning SVMA likelihood evaluation, reweighting, the Hamiltonian Monte Carlo routine, and posterior consistency; and supplemental proofs.

Any references to equations, figures, assumptions, propositions, lemmas, or sections that are not preceded by "C." refer to the main article.

C.1 Simulation study

To illustrate the workings of the SVMA approach, I conduct a small simulation study with two observed variables and two shocks. I show that prior information about the smoothness of the IRFs can substantially sharpen posterior inference. It is thus desirable to use an approach, like the SVMA approach, for which prior information about smoothness is directly controlled. I also illustrate the consequences of misspecifying the prior.

The illustration is based on the bivariate example from Section 2 with $n = 2$ and $q = 10$, cf. Figure 1. The number of parameters is $n^2(q + 1) = 2^2(10 + 1) = 44$, smaller than the dimensionality of realistic empirical applications but sufficient to elucidate the flexibility, transparency, and effectiveness of the SVAR approach.

C.1.1 Parameters, prior, and simulation settings

I consider a single parametrization, with a prior that is correctly centered but diffuse. The sample size is $T = 200$. The true IRF parameters Θ are the noninvertible ones plotted

in Figure 1. The true shock standard deviations are $\sigma_1 = 1$ (monetary policy shock) and $\sigma_2 = 0.5$ (demand shock). I first show results for the prior specification in Figure 4 with $\rho_{ij} = 0.9$ for all (i, j). The prior is centered at the true values but it expresses significant prior uncertainty about the magnitudes of the individual impulse responses. The prior on $\sigma = (\sigma_1, \sigma_2)$ is median-unbiased for the true values but it is very diffuse, with prior standard deviation of $\log \sigma_j$ equal to $\tau_j^\sigma = 2$ for $j = 1, 2$.

I simulate a single sample of artificial data from the Gaussian SVMA model and then run the HMC algorithm using the Whittle likelihood (I do not reweight the draws as in Section C.3.2). I take 10,000 MCMC steps, storing every 10th step and discarding the first 3,000 steps as burn-in.[C.1] The full computation takes less than 3 hours in Matlab 8.6 on a personal laptop with 2.3 GHz Intel CPU and 8 GB RAM. Below I provide graphical diagnostics on the convergence and mixing of the MCMC chain.

C.1.2 Baseline results

Figure C.1 shows that the posterior for the IRFs accurately estimates the true values and that the data serves to substantially reduce the prior uncertainty. The posterior means are generally close to the truth, although the means for two of the IRFs are slightly too low in this simulation. The 5–95 percentile posterior credible intervals are mostly much narrower than the prior 90% confidence bands, so this prior specification successfully allows the researcher to learn from the data about the magnitudes of the impulse responses. Figure C.2 shows the posterior draws for the shock standard deviations and compares them with the prior distribution. The posterior draws are tightly centered around the true values despite the very diffuse prior on σ. Overall, the inference method for this choice of prior works well, despite the noninvertibility of the true IRFs.

To illustrate the importance of prior information about the smoothness of the IRFs, I run the HMC algorithm with the same specification as above, except that I set $\rho_{ij} = 0.3$ for all (i, j) in the prior, as in Figure 5. Figure C.3 summarizes the posterior distribution of the IRFs corresponding to this alternative prior. Compared to Figure C.1, the posterior credible intervals are much wider and the posterior means are less accurate estimates of the true IRFs.

The higher the degree of prior smoothness, the more do nearby impulse responses "learn from each other". Due to the prior correlation structure in Equation (7), any feature of the

[C.1] The results are virtually identical in simulations with 100,000 MCMC steps.

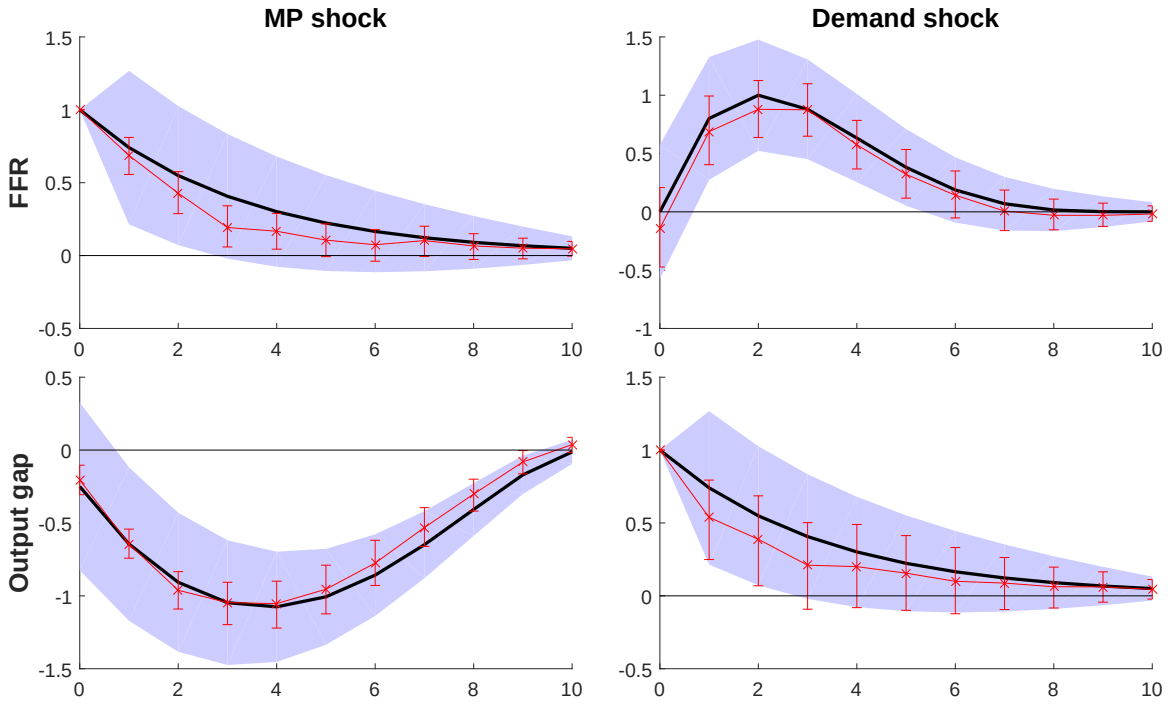

Figure C.1: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with prior smoothness $\rho_{ij} = 0.9$. The plots show true values and prior means (thick lines), prior 90% confidence bands (shaded), posterior means (crosses), and posterior 5–95 percentile intervals (vertical bars).

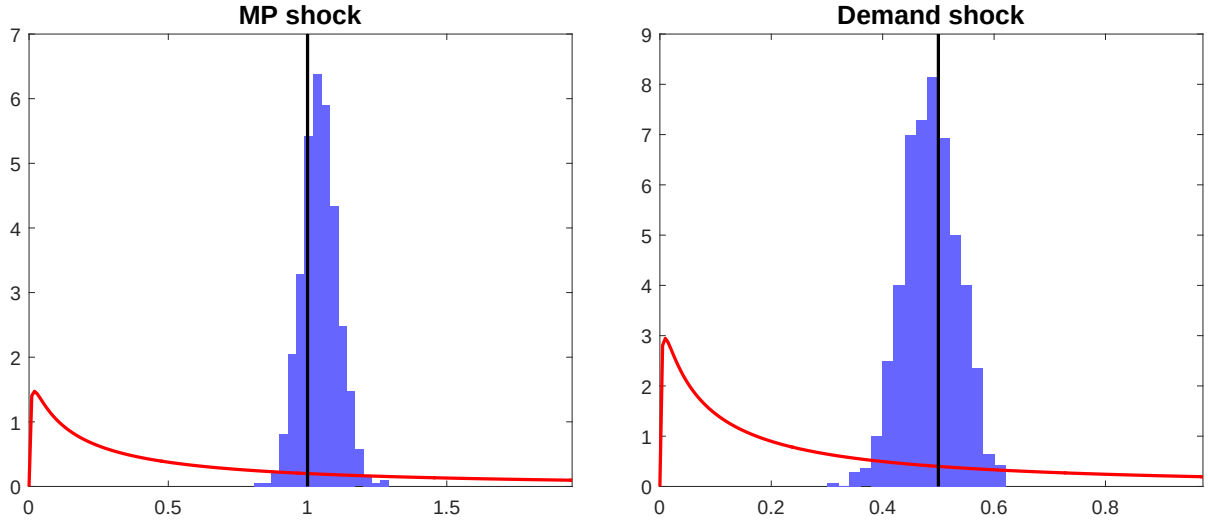

Figure C.2: Summary of posterior shock standard deviation (σ) draws for the bivariate SVMA model with prior smoothness $\rho_{ij} = 0.9$. The plots show the true value (thick vertical line), prior density (curve), and histogram of posterior draws, for each σ_j, $j = 1, 2$.

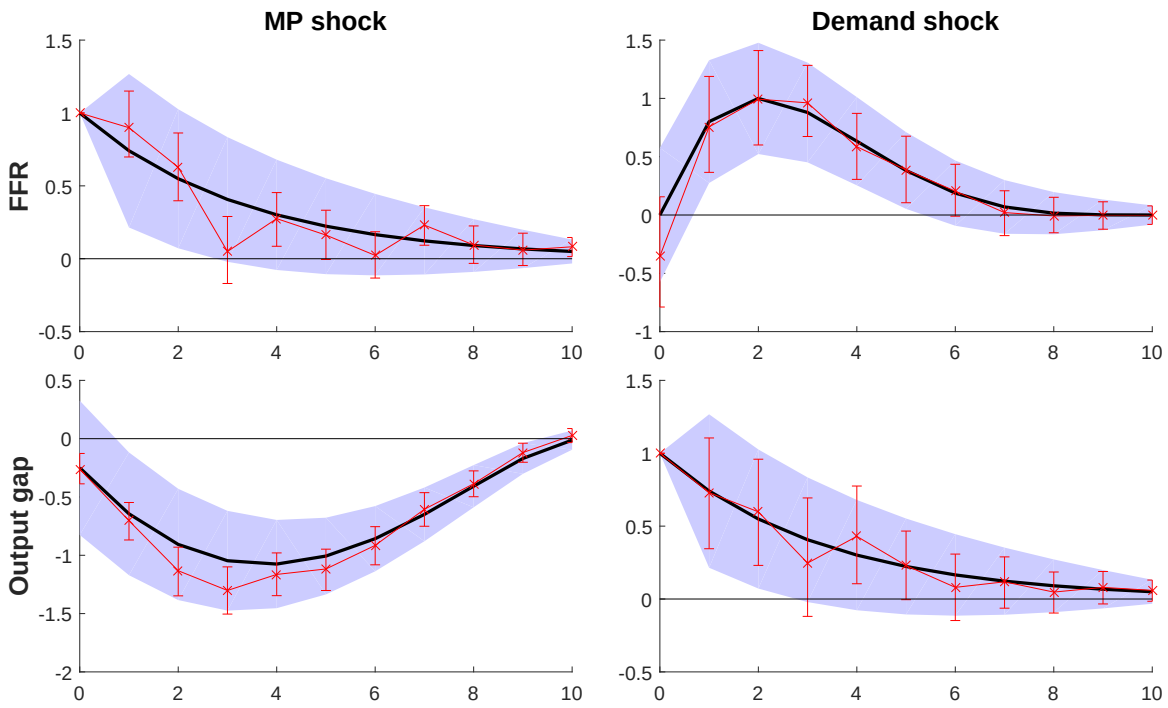

Figure C.3: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with prior smoothness $\rho_{ij} = 0.3$. See caption for Figure C.1.

data that is informative about the impulse response $\Theta_{ij,\ell}$ is also informative about $\Theta_{ij,\ell+k}$; more so for smaller values of $|k|$, and more so for larger values of the smoothness hyperparameter ρ_{ij}. Hence, a higher degree of prior smoothness reduces the effective number of free parameters in the model. If the true IRFs are not smooth but the prior imposes a lot of smoothness, posterior inference can be very inaccurate. It is therefore important to use a framework, like the SVMA approach, where prior smoothness is naturally parametrized and directly controlled. SVAR IRFs also impose smoothness *a priori*, but the degree of smoothness is implicitly controlled by the VAR lag length and restrictions on the VAR coefficients.

C.1.3 Misspecified priors

I now report results for modifications of the baseline simulation above, maintaining the prior distribution but substantially modifying the true IRFs. I maintain the same prior on IRFs and shock standard deviations as the $\rho_{ij} = 0.9$ prior. Here, however, I modify the true values of the IRFs so they no longer coincide with the prior means. I consider two such experiments: one in which the shocks have less persistent effects than the prior indicates, and one in which the true IRF of the output gap to a monetary policy shock is zero everywhere. In both

cases, the inaccurate prior is overruled by the data, delivering reasonably accurate posterior inference. This happens because the implied prior distribution of the ACF is inconsistent with the true ACF. Since the data is informative about the latter, the posterior distribution puts more weight than the prior on parameters that are consistent with the true ACF, as shown formally in Section 5.3.[C.2]

I first consider an experiment in which the prior overstates the persistence of the shock effects, i.e., the true IRFs die out quicker than indicated by the prior means $\mu_{ij,\ell}$ in Figure 4. The true IRFs are set to $\Theta_{ij,\ell} = c_{ij} e^{-0.25\ell} \mu_{ij,\ell}$ for all (i, j, ℓ), where $c_{ij} > 0$ is chosen so that $\max_\ell |\Theta_{ij,\ell}| = \max_\ell |\mu_{ij,\ell}|$ for each IRF. The true shock standard deviations, the prior ($\rho_{ij} = 0.9$), the sample size, and the HMC settings are exactly as in Section C.1.2. Figure C.4 compares these true IRFs to the prior distribution. The figure also summarizes the posterior distribution for the IRFs. The posterior is not perfectly centered but is much closer to the truth than the prior is. Figure C.5 shows why this is the case: The prior distribution on (Θ, σ) implies a distribution for auto- and cross-correlations of observed variables that is at odds with the true ACF. Since the data is informative about the ACF, the posterior distribution for IRFs puts higher weight than the prior on IRFs that are consistent with the true auto- and cross-correlations.

The second experiment considers a prior that misspecifies the cross-correlations between the observed variables. I set the true IRFs equal to the prior means in Figure 4, except that the true IRF of the output gap to a monetary policy shock equals zero, i.e., $\Theta_{21,\ell} = 0$ for $0 \leq \ell \leq q$. The true shock standard deviations, the prior ($\rho_{ij} = 0.9$), the sample size, and the HMC settings are as above. Figure C.6 shows that posterior inference is accurate despite the misspecified prior. Again, Figure C.7 demonstrates how the data corrects the prior distribution on auto- and cross-correlations, thus pulling the posterior on IRFs toward the true values (although here the true ACF is not estimated as accurately as in Figure C.5).

[C.2] By the same token, if the true parameter values were chosen to be observationally equivalent to the prior medians in Figure 4 (i.e., they imply the same ACF), then the posterior would look the same as in Figures C.1 and C.2 up to simulation noise, even though the true parameters could be very different from the prior medians. Hence, not all misspecified priors can be corrected by the data, cf. Section 5.3.

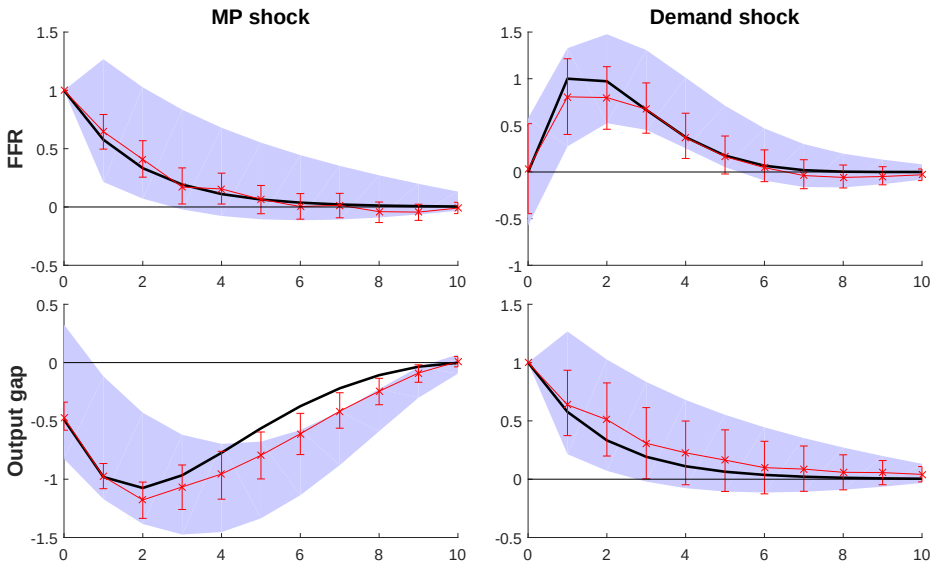

Figure C.4: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with a prior that is too persistent relative to the true parameter values. The plots show true values (thick lines), prior 90% confidence bands (shaded), posterior means (crosses), and posterior 5–95 percentile intervals (vertical bars). The prior means (not shown) are the midpoints of the prior confidence bands, as in Figure 4.

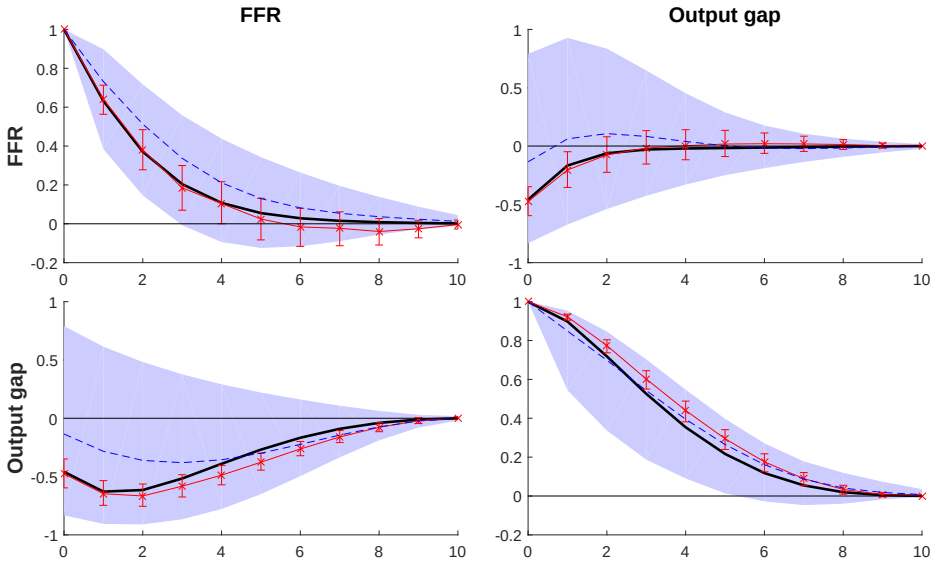

Figure C.5: Posterior auto- and cross-correlation draws for the bivariate SVMA model with a prior that misspecifices the persistence of the IRFs. The displays plot draws of $\text{Corr}(y_{i,t}, y_{j,t-k} \mid \Theta, \sigma)$, where i indexes rows, j indexes columns, and k runs along the horizontal axes. The top right display, say, concerns cross-correlations between the FFR and lags of the output gap. The plots show true values (thick lines), prior means (dashed lines) and 5–95 percentile confidence bands (shaded), and posterior means (crosses) and 5–95 percentile intervals (vertical bars).

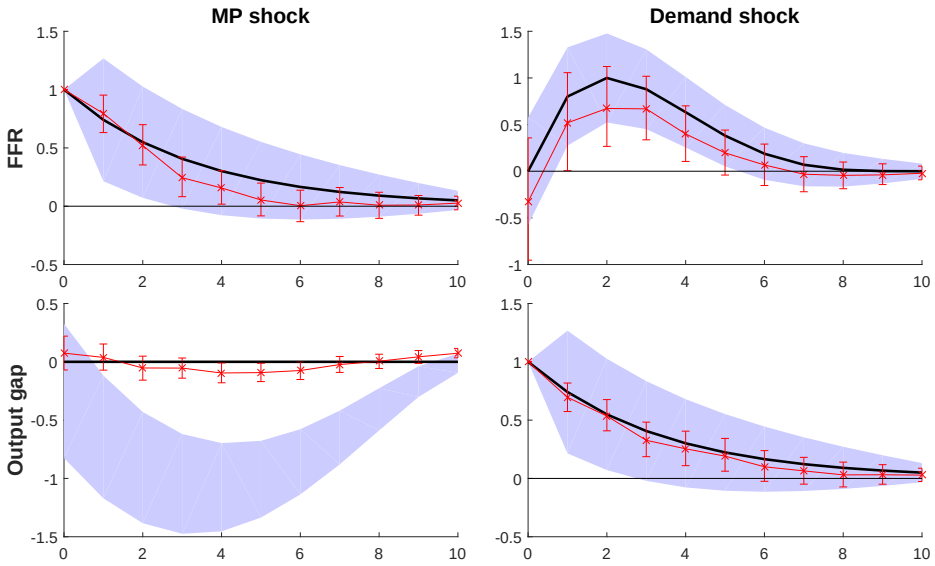

Figure C.6: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with a prior that misspecifies the cross-correlations between variables. See caption for Figure C.4.

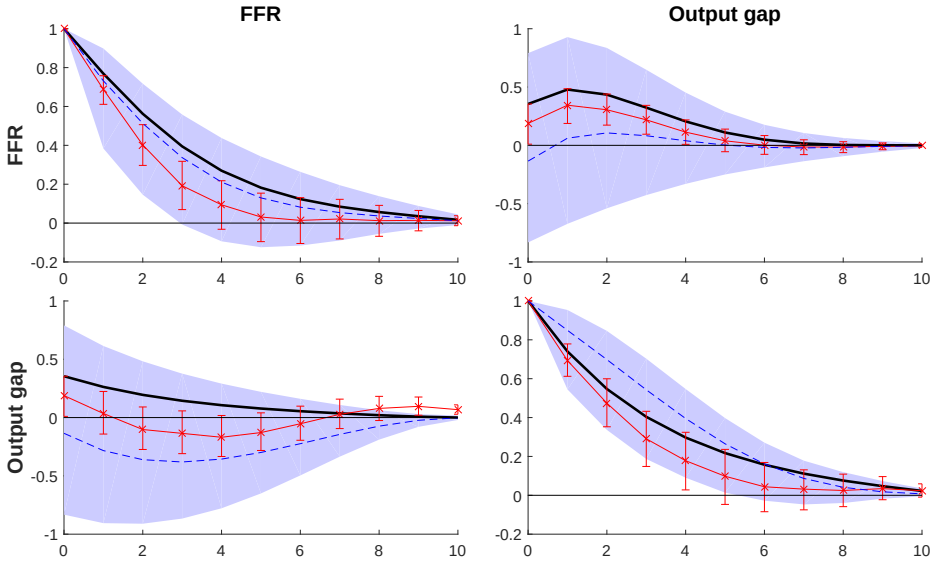

Figure C.7: Posterior autocorrelation draws for the bivariate SVMA model with a prior that misspecifies the cross-correlations between variables. See caption for Figure C.5.

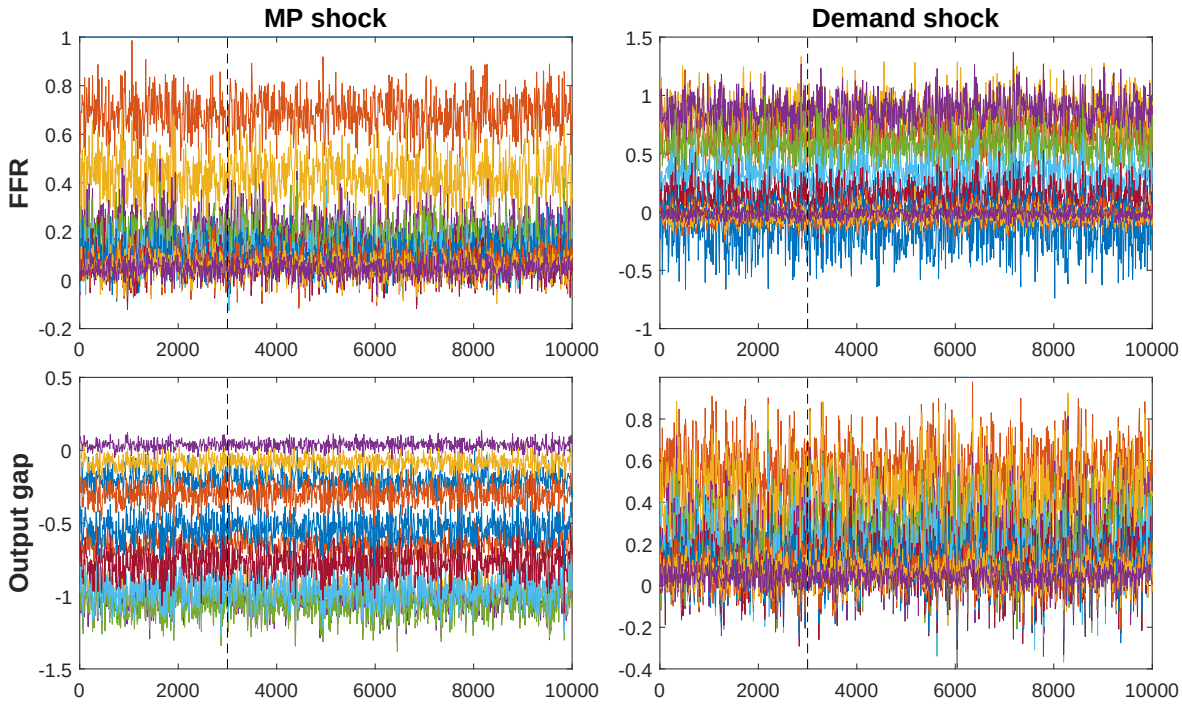

Figure C.8: MCMC chains for each IRF parameter (Θ) in the $\rho_{ij} = 0.9$ simulations in Section C.1.2. Each jagged line represents a different impulse response parameter (two of them are normalized at 1). The vertical dashed line marks the burn-in time, before which all draws are discarded. The horizontal axes are in units of MCMC steps, not stored draws (every 10th step is stored).

C.1.4 MCMC diagnostics

I report diagnostics for the baseline $\rho_{ij} = 0.9$ bivariate simulation in Section C.1.2, but diagnostics for other specifications in this paper are similar. The average HMC acceptance rate is slightly higher than 0.60, which is the rate targeted by the NUTS algorithm when tuning the HMC step size. The score of the posterior was evaluated about 382,000 times. Figures C.8 and C.9 show the MCMC chains for the IRF and log shock standard deviation draws. Figures C.10 and C.11 show the autocorrelation functions of the draws.

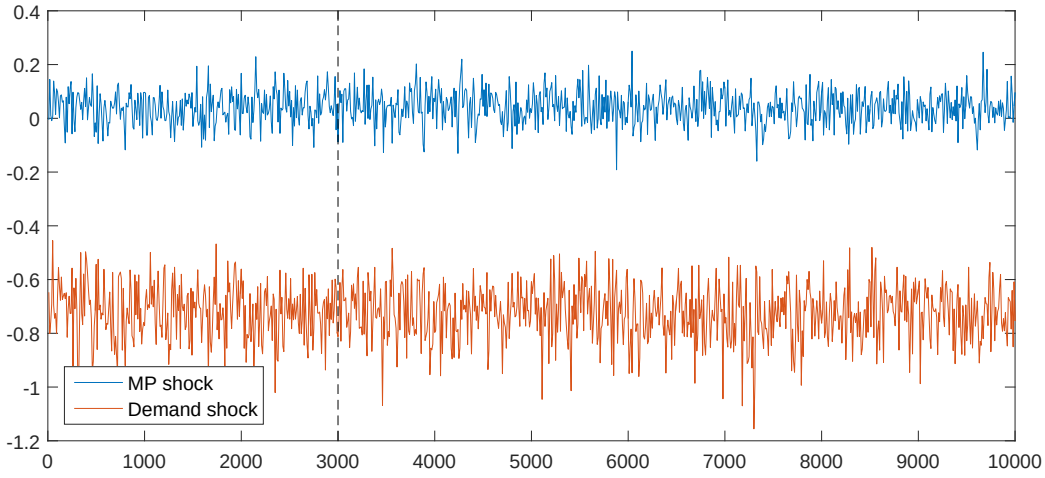

Figure C.9: MCMC chains for each log shock standard deviation parameter ($\log \sigma$) in the $\rho_{ij} = 0.9$ simulations in Section C.1.2. See caption for Figure C.8.

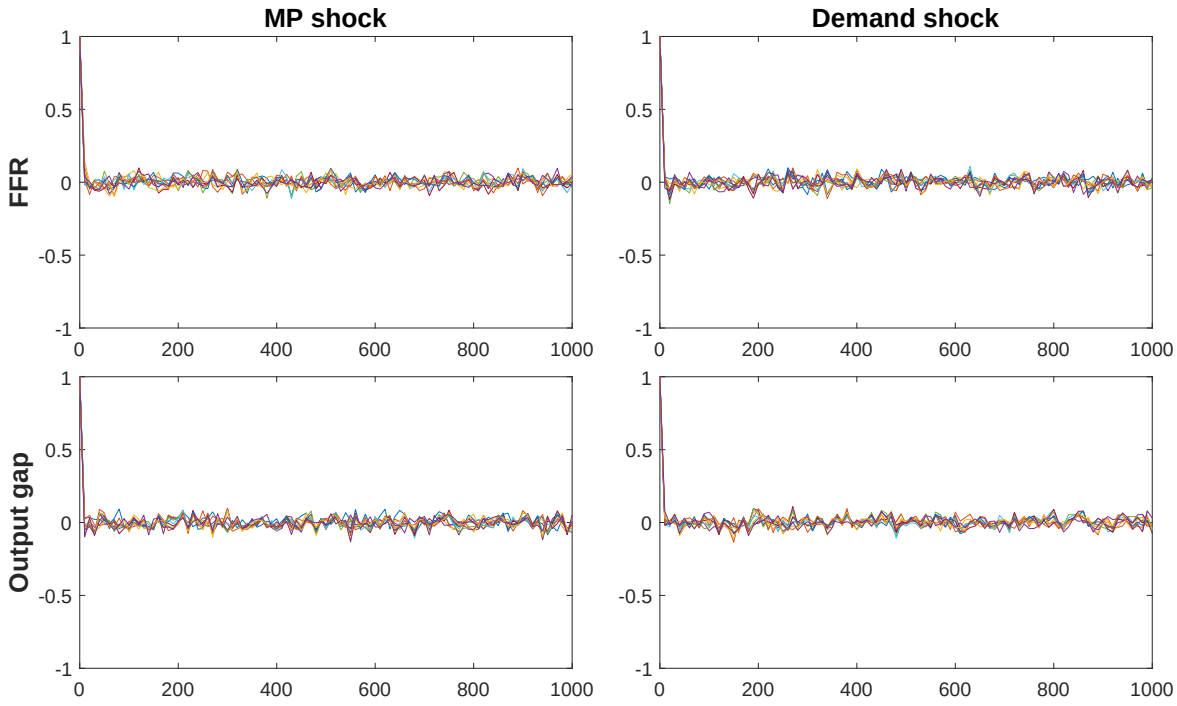

Figure C.10: Autocorrelation functions for HMC draws of each IRF parameter (Θ) in the $\rho_{ij} = 0.9$ simulations in Section C.1.2. Each jagged line represents a different impulse response parameter. Only draws after burn-in were used to computed these figures. The autocorrelation lag is shown on the horizontal axes in units of MCMC steps.

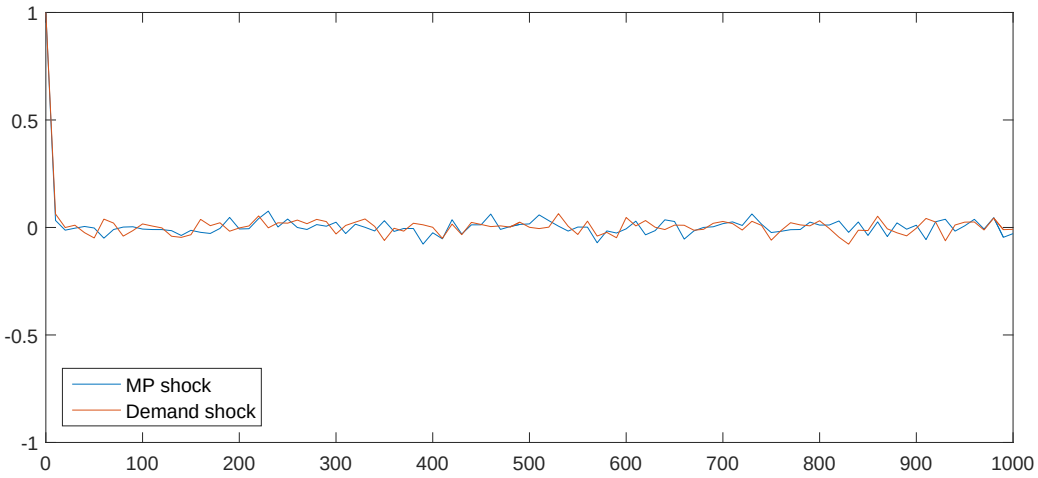

Figure C.11: Autocorrelation functions for HMC draws of each log shock standard deviation parameter ($\log \sigma$) in the $\rho_{ij} = 0.9$ simulations in Section C.1.2. See caption for Figure C.10.

C.2 Application: Additional results and discussion

This subsection presents additional details, results, and discussion related to the news shock application in Section 4. First, I detail the data construction. Second, I summarize the posteriors for long-run (cumulative) impulse responses. Third, I demonstrate the economic importance of noninvertibility. Fourth, I use the Kalman smoother to draw inference about the shocks. Fifth, I examine the sensitivity of posterior inference to the choice of prior. Sixth, I assess the model's fit. Seventh, I compare my empirical results to the literature. Finally, I show that the SVMA procedure accurately estimates IRFs on simulated data.

C.2.1 Data construction

TFP growth equals 100 times the log growth rate of TFP and is taken from the data appendix to Fernald (2014).[C.3] The remaining data is from the St. Louis Federal Reserve's FRED database.[C.4] Real GDP growth is given by 100 times the log growth rate of seasonally adjusted GDP per capita in chained dollars, as measured by the Bureau of Economic Analysis (NIPA Table 7.1, line 10). My real interest rate series equals the nominal policy interest rate minus the contemporaneous inflation rate.[C.5] The nominal policy rate is the average effective federal funds rate, expressed as a quarterly rate. The inflation rate equals 100 times the log growth rate in the seasonally adjusted implicit price deflator for the non-farm business sector, as reported by the Bureau of Labor Statistics.

I detrend the three data series to remove secular level changes that are arguably unrelated to the business cycle. Following Stock & Watson (2012), I estimate the trend in each series using a biweight kernel smoother with a bandwidth of 100 quarters; the trends are then subtracted from the raw series. The data is plotted in the Online Appendix.

Figure C.12 plots the raw data and estimated trends.

[C.3] The TFP measure is based on a growth accounting method that adjusts for differing marginal products of capital across sectors as well as changes over time in labor quality and labor's share of income. Fernald (2014) also estimates utilization-adjusted TFP, but the adjustment is model-based and reliant on estimates from annual regressions on a separate dataset, so I prefer the simpler series. Data downloaded July 14, 2015.

[C.4] FRED series codes: A939RX0Q048SBEA (real GDP per capita), FEDFUNDS (effective federal funds rate), and IPDNBS (implicit price deflator, non-farm business sector). Data downloaded August 13, 2015.

[C.5] If agents form inflation expectations under the presumption that quarterly inflation follows a random walk, then my measure of the real interest rate equals the conventional *ex ante* real interest rate.

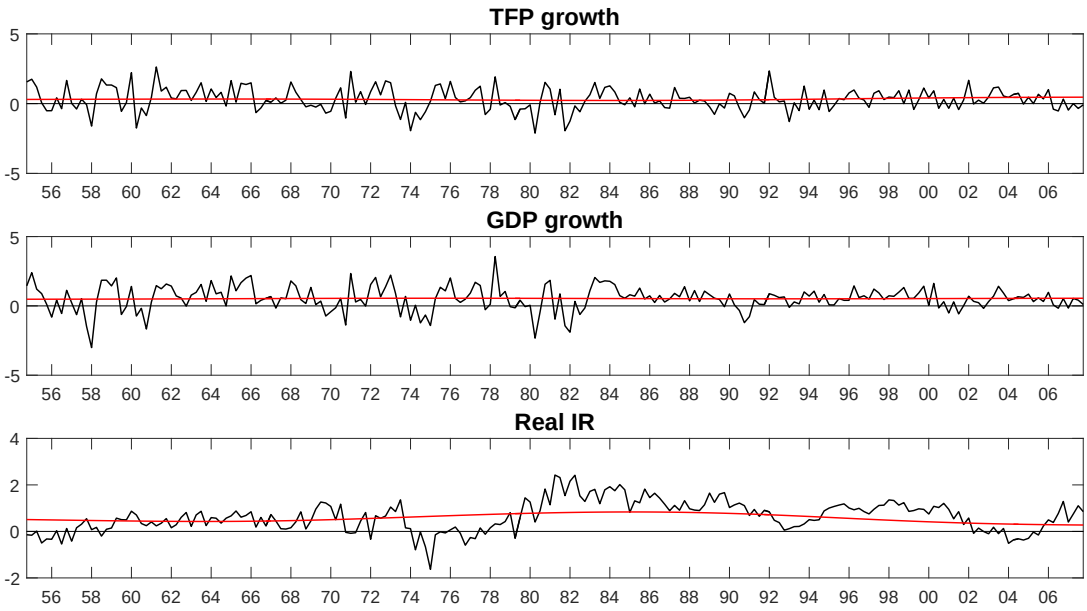

Figure C.12: Raw data on TFP growth, GDP growth, and the real interest rate (IR), along with estimated time-varying trends (smooth curves). The final data used in the empirical analysis are differences between the raw series and the trends.

C.2.2 Long-run responses

Figure C.13 plots the posterior distribution of long-run (i.e., cumulative) impulse responses $\sum_{\ell=0}^{q} \Theta_{ij,\ell}$ for each variable-shock combination (i, j).

C.2.3 Economic significance of noninvertibility

The noninvertibility of the estimated IRFs is economically significant. Figure C.14 summarizes the posterior distribution of those invertible IRFs that are closest to the actual (possibly noninvertible) IRFs. Specifically, for each posterior draw (Θ, σ) I compute the parameter vector $(\tilde{\Theta}, \tilde{\sigma})$ that minimizes the Frobenius distance $\|\Theta \operatorname{diag}(\sigma) - \tilde{\Theta} \operatorname{diag}(\tilde{\sigma})\|$ over parameters for which $\tilde{\Theta}$ is invertible and $(\tilde{\Theta}, \tilde{\sigma})$ generates the same ACF as (Θ, σ).[C.6] While the invertible IRFs for the productivity and monetary policy shocks are similar to the unrestricted IRFs,

[C.6] According to Appendix A.2, $(\tilde{\Theta}, \tilde{\sigma})$ is obtained as follows. First apply transformation (ii) in Proposition 3 several times to (Θ, σ) in order to flip all roots outside the unit circle. Denote the resulting invertible parameters by $(\check{\Theta}, \check{\sigma})$. Then $\tilde{\Theta} \operatorname{diag}(\tilde{\sigma}) = \check{\Theta} \operatorname{diag}(\check{\sigma}) Q$, where Q is the orthogonal matrix that minimizes $\|\Theta \operatorname{diag}(\sigma) - \check{\Theta} \operatorname{diag}(\check{\sigma}) Q\|$. This is an "orthogonal Procrustes problem", whose solution is well known.

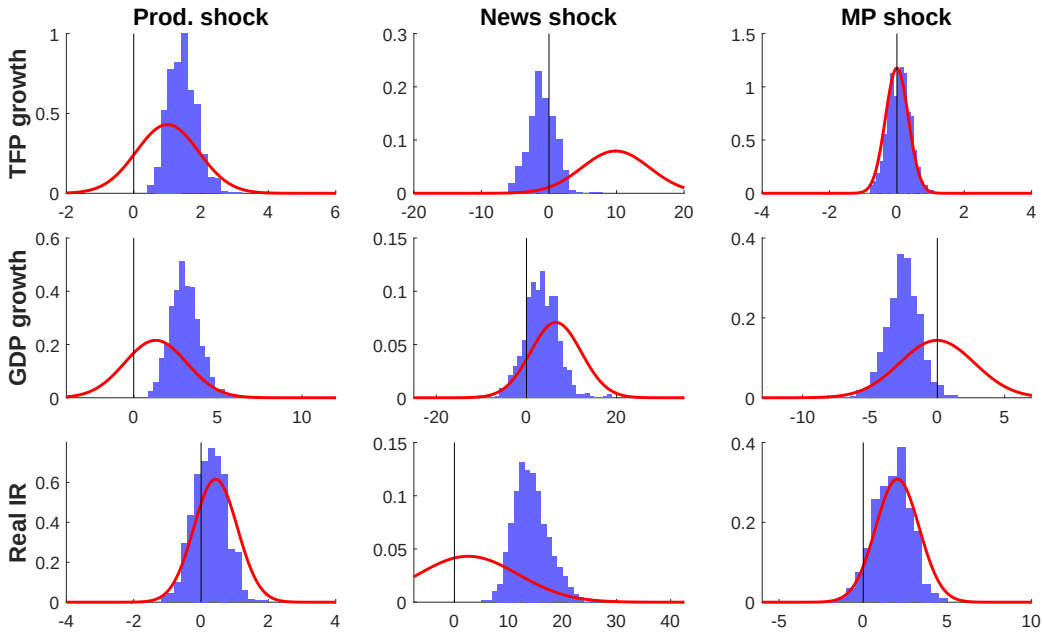

Figure C.13: Histograms of posterior draws of long-run impulse responses $\sum_{\ell=0}^{q} \Theta_{ij,\ell}$ for each (i, j), news shock application. Curves are prior densities. Histograms and curves each integrate to 1.

the invertible news shock IRFs look nothing like the actual estimated IRFs.[C.7] Thus, no SVAR identification scheme can deliver accurate inference about the effects of technological news shocks in this dataset.

C.2.4 Inference about shocks

Figure C.15 shows the time series of posterior means for the structural shocks. For each posterior draw of the structural parameters (Θ, σ), I compute $E(\varepsilon_t \mid \Theta, \sigma, Y_T)$ using the smoothing recursions corresponding to the Gaussian state-space representation in Section C.3.1 (Durbin & Koopman, 2012, p. 157), and then I average over draws. If the structural shocks are in fact non-Gaussian, the smoother still delivers mean-square-error-optimal linear estimates of the shocks. If desired, draws from the full joint posterior distribution of the shocks can be obtained from a simulation smoother (Durbin & Koopman, 2012, Ch. 4.9). It is also straight-forward to draw from the predictive distribution of future values of the data using standard methods for state-space models.

[C.7] Figure C.14 cannot be interpreted as the posterior distribution corresponding to a prior which truncates the prior from Figure 6 to the invertible region. It is difficult to sample from this truncated posterior, as essentially none of the unrestricted posterior draws are invertible, so an accept-reject scheme is inapplicable.

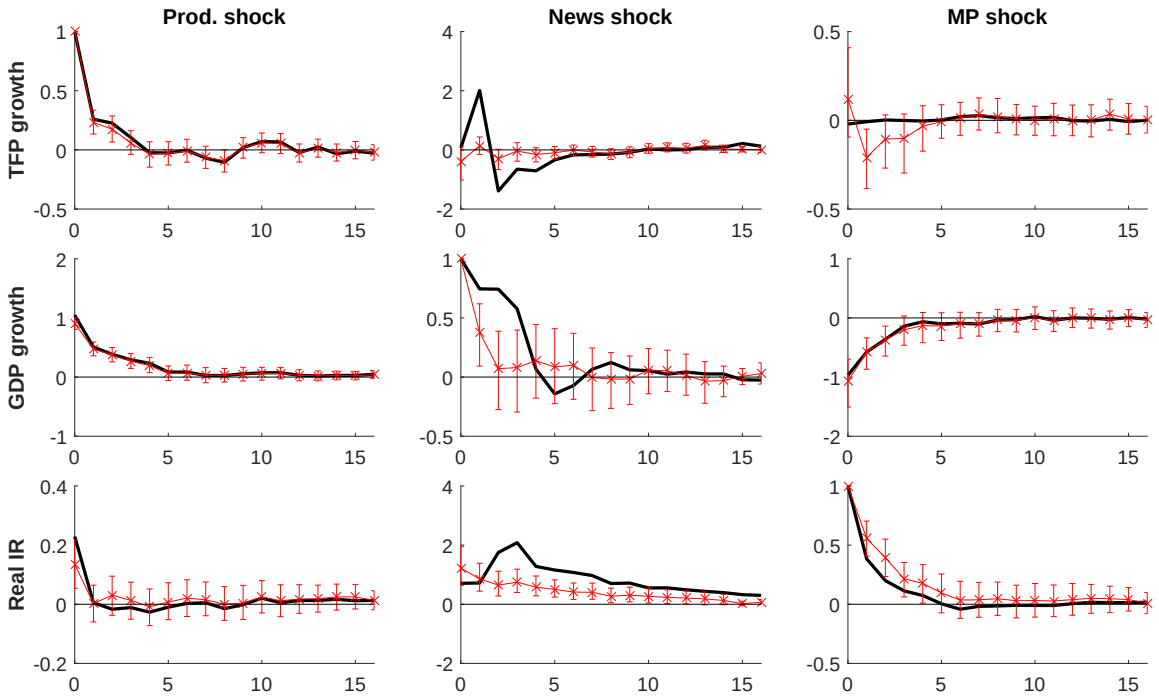

Figure C.14: Posterior distribution of the invertible IRFs that are closest to the actual IRFs, news shock application. The figure shows posterior means of actual IRFs from Figure 7 (thick lines), posterior means of the closest invertible IRFs (crosses), and posterior 5–95 percentile intervals for these invertible IRFs (vertical bars).

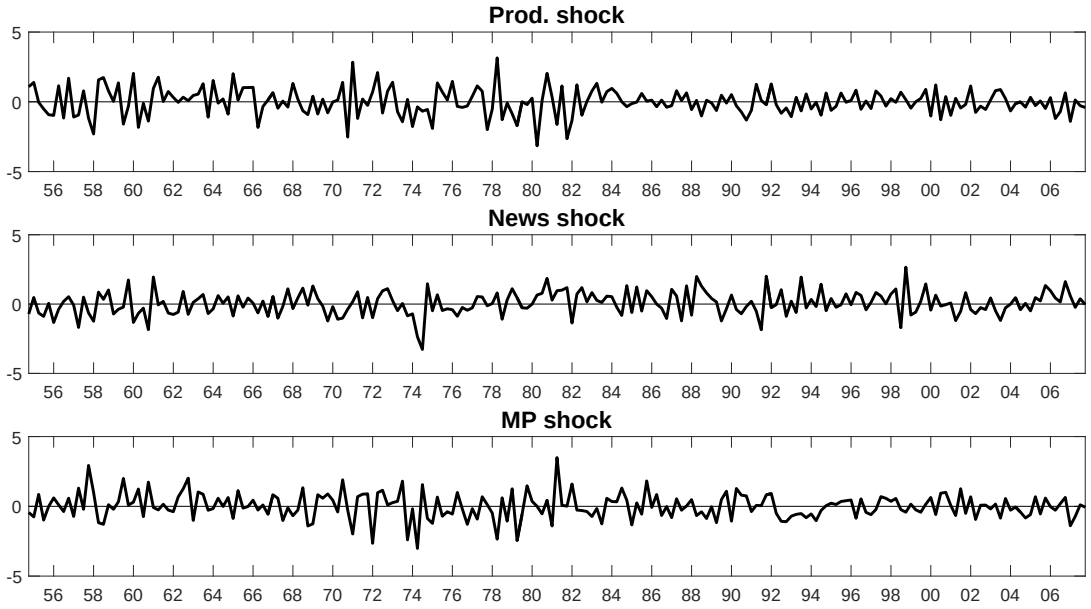

Figure C.15: Posterior means of *standardized* structural shocks ($\varepsilon_{jt}/\sigma_j$) at each point in time, news shock application.

C.2.5 Prior sensitivity

To gauge the robustness of posterior inference with respect to the choice of prior, I compute the sensitivity measure "PS" of Müller (2012). This measure captures the first-order approximate effect on the posterior means of changing the prior mean hyperparameters. Let θ denote the vector containing all impulse responses and log shock standard deviations of the SVMA model, and let e_k denote the k-th unit vector. Because my prior for θ is a member of an exponential family, the Müller (2012) PS measure for parameter θ_k equals

$$PS_k = \max_{\nu:\, \sqrt{\nu' \mathrm{Var}(\theta)^{-1} \nu} = 1} \frac{\partial E(\theta_k \mid Y_T)}{\partial E(\theta)'} \nu = \sqrt{e_k' \mathrm{Var}(\theta \mid Y_T) \mathrm{Var}(\theta)^{-1} \mathrm{Var}(\theta \mid Y_T) e_k}. \tag{C.1}$$

This is the largest (local) change that can be induced in the posterior mean of θ_k from changing the prior means of the components of θ by the multivariate equivalent of 1 prior standard deviation.[C.8] PS_k depends only on the prior and posterior variance matrices $\mathrm{Var}(\theta)$ and $\mathrm{Var}(\theta \mid Y_T)$, which are easily obtained from the HMC output.

Figure C.16 plots the posterior means of the impulse responses along with $\pm PS_k$ intervals (where the index k corresponds to the (i, j, ℓ) combination for each impulse response). The wider the band around an impulse response, the more sensitive is the posterior mean of that impulse response to (local) changes in the prior. In economic terms, most of the posterior means are seen to be insensitive to changes in the prior means of magnitudes smaller than 1 prior standard deviation. The most prior-sensitive posterior inferences, economically speaking, concern the IRF of GDP growth to a news shock, but large changes in the prior means are necessary to alter the qualitative features of the posterior mean IRF.

C.2.6 Posterior predictive analysis

I conduct a posterior predictive analysis to identify ways to improve the fit of the Gaussian SVMA model (Geweke, 2010, Ch. 2.4.2). For each posterior parameter draw produced by HMC, I simulate an artificial dataset of sample size $T = 213$ from a Gaussian SVMA model with the given parameters. On each artificial dataset I compute four checking functions. First and second, the skewness and excess kurtosis of each series. Third, the long-run autocorrelation of each series, defined as the Newey-West long-run variance estimator (20 lags) divided by the sample variance. Fourth, I run a reduced-form VAR regression of the

[C.8] In particular, $PS_k \geq \max_b |\partial E(\theta_k \mid Y_T)/\partial E(\theta_b)| \sqrt{\mathrm{Var}(\theta_b)}$. Whereas PS_k is a local measure, the effects of large changes in the prior can be evaluated using reweighting (Lopes & Tobias, 2011, Sec. 2.4).

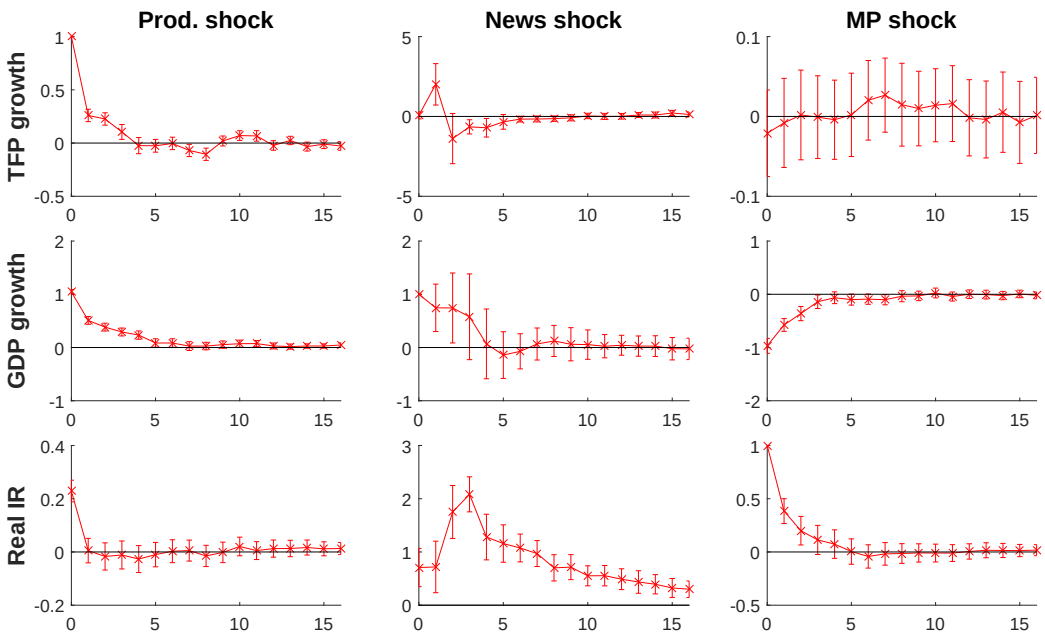

Figure C.16: PS_k measure of the sensitivity of the posterior IRF means with respect to changes in the prior means of all parameters, cf. (C.1), in the news shock application. The symmetric vertical bars have length $2PS_k$ and are centered around the corresponding posterior means (crosses).

three-dimensional data vector y_t on its 8 first lags and a constant; then I compute the first autocorrelation of the squared VAR residuals for each of the three series. The third measure captures persistence, while the fourth measure captures volatility clustering in forecast errors.

Figure C.17 shows the distribution of checking function values across simulated datasets, as well as the corresponding checking function values for the actual data. The Gaussian SVMA model does not capture the skewness and kurtosis of GDP growth; essentially, the model does not generate recessions that are sufficiently severe relative to the size of booms. The model somewhat undershoots the persistence and kurtosis of the real interest rate. The fourth column suggests that forecast errors for TFP and GDP growth exhibit volatility clustering in the data, which is not captured by the Gaussian SVMA model.

The results point to three fruitful model extensions. First, introducing stochastic volatility in the SVMA model would allow for better fit along the dimensions of kurtosis and forecast error volatility clustering. Second, nonlinearities or skewed shocks could capture the negative skewness of GDP growth. Finally, increasing the MA lag length q would allow the model to better capture the persistence of the real interest rate, although this is not a major concern, as I am primarily interested in shorter-run impulse responses.

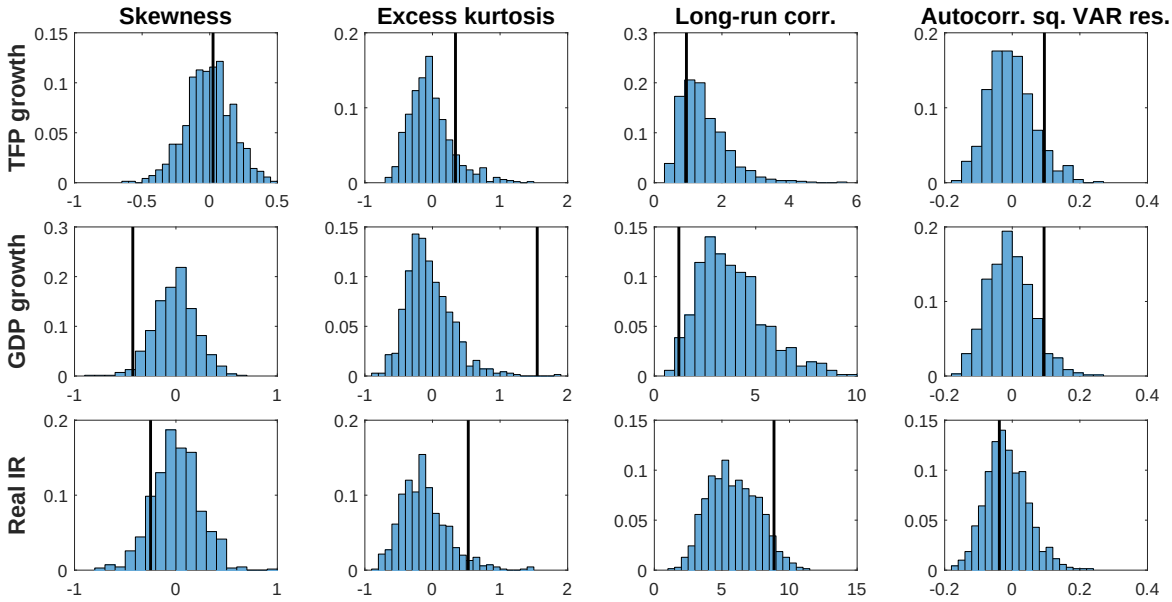

Figure C.17: Posterior predictive checks, news shock application. Observed variables along rows, checking functions along columns. Histograms show the distribution of checking function values on simulated datasets based on the posterior parameter draws; thick vertical lines mark checking function values on actual data. Checking functions from left to right: skewness; excess kurtosis; Newey-West long-run variance estimate (20 lags) divided by sample variance; first autocorrelation of squared residuals from a VAR regression of y_t on a constant and 8 lags.

C.2.7 Comparison with the literature

My conclusion that technological news shocks are not important for explaining business cycles is consistent with the literature, but my method is the first to allow for noninvertibility without additional assumptions. Forni, Gambetti & Sala (2014) estimate small effects of technological news shocks in a factor-augmented SVAR. Their empirical strategy may overcome the noninvertibility issue if technological news are well captured by the first few principal components of their large macroeconomic panel data set. They confirm that low-dimensional systems (without factors) are noninvertible. Papers that estimate fully-specified DSGE models with news shocks also tend to find a limited role for technological news, cf. the review by Beaudry & Portier (2014, Sec. 4.2.2). Unlike these papers, I do not dogmatically impose restrictions implied by a particular structural model.

Several SVAR papers on news shocks have used stock market data in an attempt to overcome the invertibility problem, cf. Beaudry & Portier (2014, Sec. 3). Such SVAR specifications may be valid if the stock market is a good proxy for the news shock, i.e., if the market responds immediately and forcefully upon arrival of technological news. On the other hand, if market movements are highly contaminated by other types of shocks, incorporating stock market data may lead to biased SVAR estimates.

C.2.8 Consistency check with simulated data

I show that the SVMA approach, with the same prior and HMC settings as in Section 4, can recover the true IRFs when applied to data generated by the log-linearized Sims (2012) DSGE model. I simulate data for the three observed variables from an SVMA model with i.i.d. Gaussian shocks. The true IRFs are those implied by the log-linearized Sims (2012) model (baseline calibration) out to horizon $q = 16$, yielding a noninvertible representation. The true shock standard deviations are set to $\sigma = (0.5, 0.5, 0.5)'$. Note that the prior for the IRF of TFP growth to the news shock is not centered at the true IRF, as explained in Section 4. The sample size is the same as for the actual data ($T = 213$).

Figures C.18 and C.19 summarize the posterior draws produced by the HMC algorithm when applied to the simulated data set. The posterior means accurately locate the true parameter values. The equal-tailed 90% posterior credible intervals are tightly concentrated around the truth in most cases. In particular, inference about the shock standard deviation parameters is precise despite the very diffuse prior.

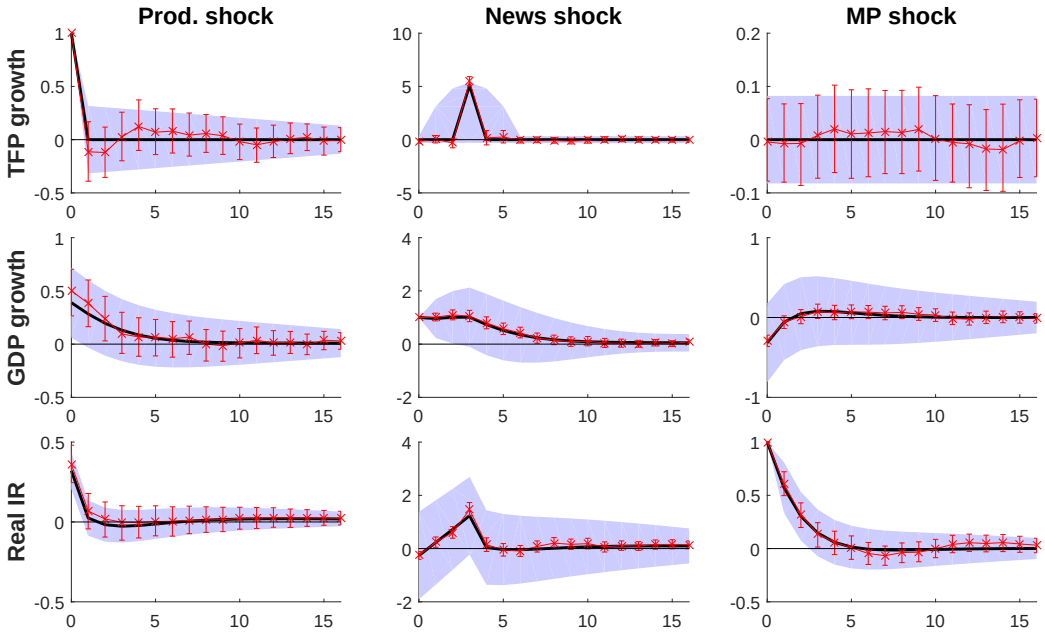

Figure C.18: Summary of posterior IRF (Θ) draws, simulated news shock data. See caption for Figure C.1.

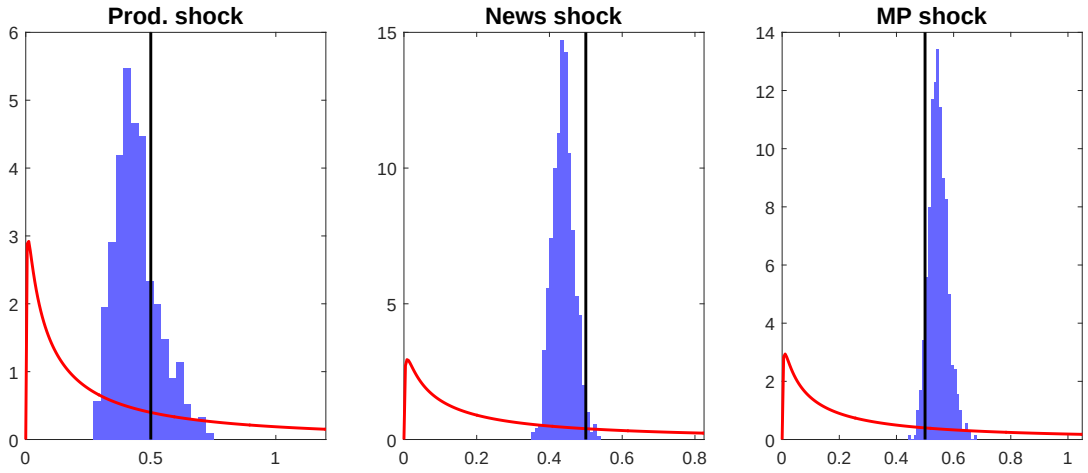

Figure C.19: Summary of posterior shock standard deviation (σ) draws, simulated news shock data. See caption for Figure C.2.

C.3 Exact likelihood and reweighting

I now describe the Kalman filter for SVMA models and the reweighting procedure for translating Whittle posterior parameter draws into draws from the exact Gaussian posterior.

C.3.1 Kalman filter

The state space representation of the SVMA model is

$$y_{i,t} = \Psi_i \alpha_t, \quad i = 1, \ldots, n,\ t = 1, \ldots, T,$$

$$\alpha_t = \begin{pmatrix} 0 & 0 \\ I_{nq} & 0 \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} \tilde{\varepsilon}_t \\ 0 \end{pmatrix}, \quad \tilde{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(0, I_n), \quad t = 2, 3, \ldots, T,$$

$$\alpha_1 \sim N(0, I_{n(q+1)}),$$

where Ψ_i is the $n(q+1)$-dimensional i-th row vector of $\Psi = (\Psi_0, \Psi_1, \ldots, \Psi_q) = \Theta \operatorname{diag}(\sigma)$, $\tilde{\varepsilon}_t$ is the n-dimensional standardized structural shock vector (each element has variance 1), and $\alpha_t = (\tilde{\varepsilon}_t', \tilde{\varepsilon}_{t-1}', \ldots, \tilde{\varepsilon}_{t-q}')'$ is the $n(q+1)$-dimensional state vector.

I use the "univariate treatment of multivariate series" Kalman filter in Durbin & Koopman (2012, Ch. 6.4), since that algorithm avoids inverting large matrices. For my purposes, the algorithm is as follows.

1. Initialize the state forecast mean $a_{1,1} = 0$ and state forecast variance $Z_{1,1} = I_{n(q+1)}$. Set $t = 1$.

2. For each $i = 1, \ldots, n$:

   (a) Compute the forecast error $v_{i,t} = y_{i,t} - \Psi_i a_{i,t}$, forecast variance $\lambda_{i,t} = \Psi_i Z_{i,t} \Psi_i'$, and Kalman gain $g_{i,t} = (1/\lambda_{i,t}) Z_{i,t} \Psi_i'$.

   (b) Compute the log likelihood contribution: $L_{i,t} = -\frac{1}{2}(\log \lambda_{i,t} + v_{i,t}^2 / \lambda_{i,t})$.

   (c) Update the state forecast mean: $a_{i+1,t} = a_{i,t} + g_{i,t} v_{i,t}$.

   (d) Update the state forecast variance: $Z_{i+1,t} = Z_{i,t} - \lambda_{i,t} g_{i,t} g_{i,t}'$.

3. Let $\tilde{a}_{n+1,t}$ denote the first nq elements of $a_{n+1,t}$, and let $\tilde{Z}_{n+1,t}$ denote the upper left $nq \times nq$ block of $Z_{n+1,t}$. Set

$$a_{1,t+1} = \begin{pmatrix} 0 \\ \tilde{a}_{n+1,t} \end{pmatrix}, \quad Z_{1,t+1} = \begin{pmatrix} I_n & 0 \\ 0 & \tilde{Z}_{n+1,t} \end{pmatrix}.$$

4. If $t = T$, stop. Otherwise, increment t by 1 and go to step 2.

The log likelihood $\log p_{Y|\Psi}(Y_T \mid \Psi)$ is given by $\sum_{t=1}^{T} \sum_{i=1}^{n} L_{i,t}$, up to a constant.

C.3.2 Reweighting

An optional reweighting step may be used to translate draws obtained from the Whittle-based HMC algorithm into draws from the exact Gaussian posterior density $p_{\Theta,\sigma|Y}(\Theta, \sigma \mid Y_T)$. The Whittle HMC algorithm yields draws $(\Theta^{(1)}, \sigma^{(1)}), \ldots, (\Theta^{(N)}, \sigma^{(N)})$ (after discarding a burn-in sample) from the Whittle posterior density $p^W_{\Theta,\sigma|Y}(\Theta, \sigma \mid Y_T)$. If desired, apply the following reweighting procedure to the Whittle draws:

1. For each Whittle draw $k = 1, 2, \ldots, N$, compute the relative likelihood weight

$$w_k = \frac{p_{Y|\Psi}(Y_T \mid \Psi(\Theta^{(k)}, \sigma^{(k)}))}{p^W_{Y|\Psi}(Y_T \mid \Psi(\Theta^{(k)}, \sigma^{(k)}))} \propto \frac{p_{\Theta,\sigma|Y}(\Theta^{(k)}, \sigma^{(k)} \mid Y_T)}{p^W_{\Theta,\sigma|Y}(\Theta^{(k)}, \sigma^{(k)} \mid Y_T)}.$$

2. Compute normalized weights $\tilde{w}_k = w_k / \sum_{b=1}^{N} w_b$, $k = 1, \ldots, N$.

3. Draw N samples $(\tilde{\Theta}^{(1)}, \tilde{\sigma}^{(1)}), \ldots, (\tilde{\Theta}^{(N)}, \tilde{\sigma}^{(N)})$ from the multinomial distribution with mass points $(\Theta^{(1)}, \sigma^{(1)}), \ldots, (\Theta^{(N)}, \sigma^{(N)})$ and corresponding probabilities $\tilde{w}_1, \ldots, \tilde{w}_N$.

Then $(\tilde{\Theta}^{(1)}, \tilde{\sigma}^{(1)}), \ldots, (\tilde{\Theta}^{(N)}, \tilde{\sigma}^{(N)})$ constitute N draws from the exact posterior distribution. This reweighting procedure is a Sampling-Importance-Resampling procedure (Rubin, 1988) that uses the Whittle posterior as a proposal distribution. The reweighting step is fast, as it only needs to compute the exact likelihood – not the score – for N different parameter values, where N is typically orders of magnitude smaller than the required number of likelihood/score evaluations during the HMC algorithm.

C.4 Hamiltonian Monte Carlo implementation

I here describe my implementation of the posterior simulation algorithm. First I outline my method for obtaining an initial value. Then I discuss the modifications I make to the Hoffman & Gelman (2014) algorithm. The calculations below require evaluation of the log prior density, its gradient, the log likelihood, and the score. Evaluation of the multivariate Gaussian log prior and its gradient is straight-forward; this is also the case for many other choices of priors. Evaluation of the Whittle likelihood and its score is described in Appendix A.3.

C.4.1 Initial value

The HMC algorithm produces draws from a Markov Chain whose long-run distribution is the Whittle posterior of the SVMA parameters, regardless of the initial value used for the chain. However, using an initial value near the mode of the posterior distribution can significantly speed up the convergence to the long-run distribution. I approximate the posterior mode using the following computationally cheap procedure:

1. Compute the empirical ACF of the data.
2. Run q steps of the Innovations Algorithm to obtain an invertible SVMA representation that approximately fits the empirical ACF (Brockwell & Davis, 1991, Prop. 11.4.2).[C.9] Denote these invertible parameters by $(\hat{\Theta}, \hat{\sigma})$.
3. Let $\mathcal{C}$ denote the (finite) set of complex roots of the SVMA polynomial corresponding to $(\hat{\Theta}, \hat{\sigma})$, cf. Proposition 3.
4. For each root γ_j in $\mathcal{C}$ (each complex conjugate pair of roots is treated as one root):
   (a) Let $(\check{\Theta}^{(j)}, \check{\sigma}^{(j)})$ denote the result of flipping root γ_j, i.e., of applying transformation (ii) in Proposition 3 to $(\hat{\Theta}, \hat{\sigma})$ with this root.
   (b) Determine the orthogonal matrix $Q^{(j)}$ such that $\check{\Theta}^{(j)} \operatorname{diag}(\check{\sigma})^{(j)} Q^{(j)}$ is closest to the prior mean $E(\Theta \operatorname{diag}(\sigma))$ in Frobenius norm, cf. Footnote C.6.
   (c) Obtain parameters $(\tilde{\Theta}^{(j)}, \tilde{\sigma}^{(j)})$ such that $\check{\Theta}^{(j)} \operatorname{diag}(\check{\sigma}^{(j)}) Q^{(j)} = \tilde{\Theta}^{(j)} \operatorname{diag}(\tilde{\sigma}^{(j)})$, i.e., apply transformation (a) in Proposition 3. Calculate the corresponding value of the prior density $\pi(\tilde{\Theta}^{(j)}, \tilde{\sigma}^{(j)})$.
5. Let $\tilde{j} = \operatorname{argmax}_j \pi(\tilde{\Theta}^{(j)}, \tilde{\sigma}^{(j)})$.
6. If $\pi(\tilde{\Theta}^{(\tilde{j})}, \tilde{\sigma}^{(\tilde{j})}) \leq \pi(\hat{\Theta}, \hat{\sigma})$, go to Step 7. Otherwise, set $(\hat{\Theta}, \hat{\sigma}) = (\tilde{\Theta}^{(\tilde{j})}, \tilde{\sigma}^{(\tilde{j})})$, remove γ_j (and its complex conjugate) from $\mathcal{C}$, and go back to Step 4.
7. Let the initial value for the HMC algorithm be the parameter vector of the form $((1-x)\check{\Theta} + xE(\Theta), (1-x)\check{\sigma} + xE(\sigma))$ that maximizes the posterior density, where x ranges over the grid $\{0, 0.01, \ldots, 0.99, 1\}$, and $(E(\Theta), E(\sigma))$ is the prior mean of (Θ, σ).

[C.9] In principle, the Innovations Algorithm could be run for more than q steps, but this tends to lead to numerical instability in my trials. The output of the first q steps is sufficiently accurate in my experience.

Step 2 computes a set of invertible parameters that yields a high value of the likelihood. Steps 3–6 find a set of possibly noninvertible parameters that yields a high value of the prior density while being observationally equivalent with the parameters from Step 2 (I use a "greedy" search algorithm since it is computationally prohibitive to consider all combinations of root flips). Because Steps 2–6 lexicographically prioritize maximizing the likelihood over maximizing the prior, Step 7 allows the parameters to shrink toward the prior means.

C.4.2 Modifications to NUTS algorithm

I use the HMC variant NUTS from Hoffman & Gelman (2014), which automatically tunes the step size and trajectory length of HMC. See their paper for details on the NUTS algorithm. I downloaded the code from Hoffman's website.[C.10] I make two modifications to the basic NUTS algorithm, neither of which are essential, although they do tend to improve the mixing speed of the Markov chain in my trials: step size jittering and diagonal mass matrix adaptation. These modifications are also used in the NUTS-based statistics software Stan (Stan Development Team, 2015).

Each step I draw a new HMC step size from a uniform distribution over some interval (Neal, 2011, Sec. 5.4.2.2). The jittering is started after the stepsize has been tuned as described in Hoffman & Gelman (2014, Sec. 3.2). For the applications in this paper, the step size is chosen uniformly at random from the interval $[0.5\hat{\epsilon}, 1.5\hat{\epsilon}]$, where $\hat{\epsilon}$ is the tuned step size.

I allow for a diagonal "mass matrix", where the entries along the diagonal are estimates of the posterior standard deviations of the SVMA parameters (Neal, 2011, Sec. 5.4.2.4). I first run the NUTS algorithm for a number of steps with an identity mass matrix. Then I calculate the sample standard deviations of the parameter draws over a window of subsequent steps, after which I update the mass matrix accordingly.[C.11] I update the mass matrix twice more using windows of increasing length. Finally, I freeze the mass matrix for the remainder of the NUTS algorithm. In this paper, the mass matrix is estimated over steps 300–400, steps 401–600, and steps 601–1000, and it is fixed after step 1000.

[C.10] `http://matthewdhoffman.com`

[C.11] The sample standard deviations are partially shrunk toward 1 before updating the mass matrix.

C.5 Posterior consistency: Auxiliary lemma

Following Ghosh & Ramamoorthi (2003, Thm. 1.3.4), I give general sufficient conditions for assumption (ii) of Lemma 1. These are used in the proof of Lemma 3.

Let $\Pi_\Gamma(\cdot)$ denote the marginal prior measure for parameter Γ, with parameter space Ξ_Γ. Let $p_{Y|\Gamma}(Y_T \mid \Gamma)$ denote the (possibly misspecified) likelihood function. The posterior measure is given by

$$P_{\Gamma|Y}(\mathcal{A} \mid Y_T) = \frac{\int_{\mathcal{A}} p_{Y|\Gamma}(Y_T \mid \Gamma)\Pi_\Gamma(d\Gamma)}{\int_{\Xi_\Gamma} p_{Y|\Gamma}(Y_T \mid \Gamma)\Pi_\Gamma(d\Gamma)}$$

for measurable sets $\mathcal{A} \subset \Xi_\Gamma$.

Lemma C.1. *Define the normalized log likelihood ratio* $\hat{\phi}(\Gamma) = T^{-1} \log \frac{p_{Y|\Gamma}(Y_T|\Gamma)}{p_{Y|\Gamma}(Y_T|\Gamma_0)}$ *for all* $\Gamma \in \Xi_\Gamma$. *Assume there exist a function* $\phi\colon \Xi_\Gamma \to \mathbb{R}$, *a neighborhood* $\mathcal{K}$ *of* Γ_0 *in* Ξ_Γ, *and a scalar* $\zeta < 0$ *such that the following conditions hold.*

(i) $\sup_{\Gamma\in\mathcal{K}} |\hat{\phi}(\Gamma) - \phi(\Gamma)| \xrightarrow{p} 0$.

(ii) $\phi(\Gamma)$ *is continuous at* $\Gamma = \Gamma_0$.

(iii) $\phi(\Gamma) < 0$ *for all* $\Gamma \neq \Gamma_0$.

(iv) $\sup_{\Gamma\in\mathcal{K}^c} \hat{\phi}(\Gamma) < \zeta$ *w.p.a. 1.*

(v) Γ_0 *is in the support of* $\Pi_\Gamma(\cdot)$.

Then for any neighborhood $\mathcal{U}$ *of* Γ_0 *in* Ξ_Γ, $P_{\Gamma|Y}(\mathcal{U} \mid Y_T) \xrightarrow{p} 1$.

C.6 Supplemental proofs

C.6.1 Proof of Proposition 3

As in Lippi & Reichlin (1994, p. 311), define the rational matrix function

$$R(\gamma, z) = \begin{pmatrix} \frac{z-\gamma}{1-\bar{\gamma}z} & 0 \\ 0 & I_{n-1} \end{pmatrix}, \quad \gamma, z \in \mathbb{C}.$$

Transformation (ii) corresponds to the transformation $\check{\Psi}(z) = \Psi(z)QR(\gamma_k, z)^{-1}$ if γ_k is real. If γ_k is not real, the transformation corresponds to $\check{\Psi}(z) = \tilde{\Psi}(z)\check{Q}$, where $\tilde{\Psi}(z) = \Psi(z)QR(\gamma_k, z)^{-1}\tilde{Q}R(\overline{\gamma_k}, z)^{-1}$ and $\check{Q} = \tilde{\Psi}(0)^{-1}J$ is a unitary matrix. I proceed in three steps.

STEP 1. Consider the first claim of the proposition. Let $f(\omega;\Gamma) = (2\pi)^{-1}\sum_{k=-q}^{q}\Gamma(k)e^{-\imath k\omega}$, $\omega \in [-\pi,\pi]$, denote the spectral density matrix function associated with the ACF $\Gamma(\cdot)$. Since $\Psi(z) = \Theta(z)\,\mathrm{diag}(\sigma)$ with $(\Theta,\sigma) \in \mathcal{S}(\Gamma)$, we must have $\Psi(e^{-\imath\omega})\Psi(e^{-\imath\omega})^* = 2\pi f(\omega;\Gamma)$ for all ω by the usual formula for the spectral density of a vector MA process (Brockwell & Davis, 1991, Example 11.8.1). Because $R(\gamma,e^{-\imath\omega})R(\gamma,e^{-\imath\omega})^* = I_n$ for any (γ,ω), it is easy to verify that $\check{\Psi}(z)$ – constructed by applying transformation (i) or transformation (ii) to $\Psi(z)$ – also satisfies $\check{\Psi}(e^{-\imath\omega})\check{\Psi}(e^{-\imath\omega})^* = 2\pi f(\omega;\Gamma)$. Hence, $\check{\Psi}(z) = \sum_{\ell=0}^{q}\check{\Psi}_\ell z^\ell$ is a matrix MA polynomial satisfying $\sum_{\ell=0}^{q-k}\check{\Psi}_{\ell+k}\check{\Psi}_\ell^{\,*} = \Gamma(k)$ for all $k = 0,1,\ldots,q$. In Step 2 below I show that $\check{\Psi}(z)$ is a matrix polynomial with real coefficients. By construction of $\check{\Theta}(z) = \sum_{\ell=0}^{q}\check{\Theta}_\ell z^\ell$ and $\check{\sigma}$, we then have $\sum_{\ell=0}^{q-k}\check{\Theta}_{\ell+k}\,\mathrm{diag}(\check{\sigma})^2\check{\Theta}_\ell^{\,\prime} = \Gamma(k)$ for all $k = 0,1,\ldots,q$, so $(\check{\Theta},\check{\sigma}) \in \mathcal{S}(\Gamma)$, as claimed.

STEP 2. I now show that transformation (ii) yields a real matrix polynomial $\check{\Psi}(z)$. This fact was asserted by Lippi & Reichlin (1994, pp. 317–318). I am grateful to Professor Marco Lippi for providing me with the proof arguments for Step 2; all errors are my own.

$\check{\Psi}(z)$ is clearly real if the flipped root γ_k is real (since η and Q can be chosen to be real in this case), so consider the case where we flip a pair of complex conjugate roots γ_k and $\overline{\gamma_k}$. Recall that in this case, $\check{\Psi}(z) = \tilde{\Psi}(z)\check{Q}$, where $\tilde{\Psi}(z) = \Psi(z)QR(\gamma_k,z)^{-1}\tilde{Q}R(\overline{\gamma_k},z)^{-1}$ and $\check{Q}$ is unitary. It follows from the same arguments as in Step 1 that the complex-valued matrix polynomial $\tilde{\Psi}(z) = \sum_{\ell=0}^{q}\tilde{\Psi}_\ell z^\ell$ satisfies $\sum_{\ell=0}^{q-k}\tilde{\Psi}_{\ell+k}\tilde{\Psi}_\ell^{\,*} = \Gamma(k)$ for all $k = 0,1,\ldots,q$

Let $\bar{\tilde{\Psi}}(z) = \sum_{\ell=0}^{q}\overline{\tilde{\Psi}_\ell}z^\ell$ denote the matrix polynomial obtained by conjugating the coefficients of the polynomial $\tilde{\Psi}(z)$. By construction, the roots of $\det(\tilde{\Psi}(z))$ are real or appear as complex conjugate pairs, so $\det(\bar{\tilde{\Psi}}(z))$ has the same roots as $\det(\tilde{\Psi}(z))$. Furthermore, for $k = 0,1,\ldots,q$,

$$\sum_{\ell=0}^{q-k}\bar{\tilde{\Psi}}_{\ell+k}\bar{\tilde{\Psi}}_\ell^{\,*} = \overline{\Gamma(k)} = \Gamma(k) = \sum_{\ell=0}^{q-k}\tilde{\Psi}_{\ell+k}\tilde{\Psi}_\ell^{*}.$$

By Theorem 3(b) of Lippi & Reichlin (1994), there exists a unitary $n\times n$ matrix $\tilde{\tilde{Q}}$ such that $\bar{\tilde{\Psi}}(z) = \tilde{\Psi}(z)\tilde{\tilde{Q}}$ for $z \in \mathbb{R}$. The matrix polynomial $\tilde{\Psi}(z)\tilde{\Psi}(0)^{-1}$ then has real coefficients:[C.12] For all $z \in \mathbb{R}$,

$$\tilde{\Psi}(z)\tilde{\Psi}(0)^{-1} = \left(\tilde{\Psi}(z)\tilde{\tilde{Q}}\right)\left(\tilde{\Psi}(0)\tilde{\tilde{Q}}\right)^{-1} = \bar{\tilde{\Psi}}(z)\bar{\tilde{\Psi}}(0)^{-1} = \overline{\tilde{\Psi}(z)\tilde{\Psi}(0)^{-1}}.$$

Consequently, with the real matrix J defined as in the proposition, $\check{\Psi}(z) = \tilde{\Psi}(z)\tilde{\Psi}(0)^{-1}J$ is

[C.12] $\tilde{\Psi}(0)$ is nonsingular because $\det(\Psi(0)) \neq 0$.

a matrix polynomial with real coefficients. Note that, since $\tilde{\bar{Q}}$ is unitary, the matrix

$$\tilde{\Psi}(0)\tilde{\Psi}(0)^* = \left(\tilde{\bar{\Psi}}(0)\tilde{\bar{Q}}\right)\left(\tilde{\bar{\Psi}}(0)\tilde{\bar{Q}}\right)^* = \overline{\tilde{\Psi}(0)\tilde{\Psi}(0)^*}$$

is real, symmetric, and positive definite, so J is well-defined.

Step 3. Finally, I prove the second claim of the proposition. Suppose we have a fixed element $(\check{\Theta}, \check{\sigma})$ of the identified set that we want to end up with after transforming the initial element (Θ, σ) appropriately. Define $\check{\Psi}(z) = \check{\Theta}(z)\,\mathrm{diag}(\check{\sigma})$. Since $(\Theta, \sigma), (\check{\Theta}, \check{\sigma}) \in \mathcal{S}(\Gamma)$, the two sets of SVMA parameters correspond to the same spectral density, i.e., $\Psi(e^{-\imath\omega})\Psi(e^{-\imath\omega})^* = \check{\Psi}(e^{-\imath\omega})\check{\Psi}(e^{-\imath\omega})^*$ for all $\omega \in [-\pi, \pi]$. As in the proof of Theorem 2 in Lippi & Reichlin (1994), we can apply transformation (ii) finitely many (say, b) times to $\Psi(z)$, flipping all the roots that are inside the unit circle, thus ending up with a polynomial

$$B(z) = \Psi(z)Q_1 R(\gamma_{k_1}, z)^{-1} \cdots Q_b R(\gamma_{k_b}, z)^{-1} Q_{b+1}$$

for which all roots of $\det(B(z))$ lie on or outside the unit circle. Likewise, denote the (finitely many) roots of $\det(\check{\Psi}(z))$ by $\check{\gamma}_k$, $k = 1, 2, \ldots$, and apply to $\check{\Psi}(z)$ a finite sequence of transformation (ii) to arrive at a polynomial

$$\check{B}(z) = \check{\Psi}(z)\check{Q}_1 R(\check{\gamma}_{\check{k}_1}, z)^{-1} \cdots \check{Q}_{\check{b}} R(\check{\gamma}_{\check{k}_{\check{b}}}, z)^{-1} \check{Q}_{\check{b}+1}$$

for which all roots of $\det(\check{B}(z))$ lie on or outside the unit circle. Since $\det(B(z))$ and $\det(\check{B}(z))$ have all roots on or outside the unit circle, and we have $B(e^{-\imath\omega})B(e^{-\imath\omega})^* = \check{B}(e^{-\imath\omega})\check{B}(e^{-\imath\omega})^* = 2\pi f(\omega; \Gamma)$ for all ω, there must exist an orthogonal matrix Q such that $\check{B}(z) = B(z)Q$ (Lippi & Reichlin, 1994, p. 313; Hannan, 1970, p. 69). Thus,

$$\check{\Psi}(z) = \Psi(z)Q_1 R(\gamma_{k_1}, z)^{-1} \cdots Q_b R(\gamma_{k_b}, z)^{-1} Q_{b+1} Q \check{Q}^*_{\check{b}+1} R(\check{\gamma}_{\check{k}_{\check{b}}}, z)\check{Q}^*_{\check{b}} \cdots R(\check{\gamma}_{\check{k}_1}, z)\check{Q}^*_1,$$

and

$$\det(\check{\Psi}(z)) = \det(\Psi(z))\frac{(z - \check{\gamma}_{\check{k}_1}) \cdots (z - \check{\gamma}_{\check{k}_{\check{b}}})(1 - \overline{\gamma_{k_1}}z) \cdots (1 - \overline{\gamma_{k_b}}z)}{(z - \gamma_{k_1}) \cdots (z - \gamma_{k_b})(1 - \overline{\check{\gamma}_{\check{k}_1}}z) \cdots (1 - \overline{\check{\gamma}_{\check{k}_{\check{b}}}}z)},$$

so any root of $\det(\check{\Psi}(z))$ must either equal γ_k or it must equal $1/\overline{\gamma_k}$, where γ_k is some root of $\det(\Psi(z))$. It follows that we can apply a finite sequence of transformation (ii) (i.e., an appropriate sequence of root flips) to $\Psi(z)$ to obtain a real matrix polynomial $\tilde{\check{\Psi}}(z)$ satisfying $\det(\tilde{\check{\Psi}}(z)) = \det(\check{\Psi}(z))$ for all $z \in \mathbb{C}$. Theorem 3(b) in Lippi & Reichlin (1994) then implies

that $\check{\Psi}(z)$ can be obtained from $\tilde{\tilde{\Psi}}(z)$ through transformation (i) (i.e., an orthogonal rotation, which clearly must be real). Finally, obtain $(\check{\Theta}, \check{\sigma})$ from $\check{\Psi}(z)$ by transformation (a). □

C.6.2 Proof of Lemma 2

Suppressing the arguments (Ψ), let $L_k = \log\det(f_k) + \tilde{y}_k^* f_k^{-1} \tilde{y}_k$. Then

$$\frac{\partial L_k}{\partial (f_k')} = f_k^{-1} - f_k^{-1}\tilde{y}_k\tilde{y}_k^* f_k^{-1} = C_k.$$

Writing $f_k' = \overline{\tilde{\Psi}_k}\tilde{\Psi}_k'$, we have

$$\frac{\partial \operatorname{vec}(f_k')}{\partial \operatorname{vec}(\Psi_\ell)'} = (\tilde{\Psi}_k \otimes I_n)e^{\imath\omega_k\ell} + (I_n \otimes \overline{\tilde{\Psi}_k})K_n' e^{-\imath\omega_k\ell},$$

where K_n is the $n^2 \times n^2$ commutation matrix such that $\operatorname{vec}(B') = K_n \operatorname{vec}(B)$ for any $n \times n$ matrix B (Magnus & Neudecker, 2007, Ch. 3.7). Using $\operatorname{vec}(ABC) = (C' \otimes A)\operatorname{vec}(B)$,

$$\frac{\partial L_k}{\partial \operatorname{vec}(\Psi_\ell)'} = \frac{\partial L_k}{\partial \operatorname{vec}(f_k')'}\frac{\partial \operatorname{vec}(f_k')}{\partial \operatorname{vec}(\Psi_\ell)'} = \operatorname{vec}\left(C_k\tilde{\Psi}_k e^{\imath\omega_k\ell} + \overline{C_k^*\tilde{\Psi}_k e^{\imath\omega_k\ell}}\right)'.$$

Since $C_k^* = C_k$, we get $\partial L_k/\partial\Psi_\ell = 2\operatorname{Re}\left(C_k\tilde{\Psi}_k e^{\imath\omega_k\ell}\right)$, so

$$\begin{aligned}
\frac{\partial \log p_{Y|\Psi}^W(Y_T \mid \Psi)}{\partial \Psi_\ell} &= -\frac{1}{2}\sum_{k=0}^{T-1}\frac{\partial L_k}{\partial \Psi_\ell} \\
&= -\sum_{k=0}^{T-1}\operatorname{Re}\left(C_k\sum_{\tilde{\ell}=1}^{q+1} e^{-\imath\omega_k(\tilde{\ell}-1)}\Psi_{\tilde{\ell}-1}e^{\imath\omega_k\ell}\right) \\
&= -\sum_{\tilde{\ell}=0}^{q}\operatorname{Re}\left(\sum_{k=0}^{T-1} C_k e^{-\imath\omega_k(\tilde{\ell}-\ell)}\right)\Psi_{\tilde{\ell}}.
\end{aligned}$$

Finally, $\sum_{k=0}^{T-1} C_k e^{-\imath\omega_k(\tilde{\ell}-\ell)} = \sum_{k=0}^{T-1} C_k e^{-\imath\omega_{\tilde{\ell}-\ell}k} = \tilde{C}_{\tilde{\ell}-\ell}$ for $\tilde{\ell} \geq \ell$, and $\sum_{k=0}^{T-1} C_k e^{-\imath\omega_k(\tilde{\ell}-\ell)} = \sum_{k=0}^{T-1} C_k e^{-\imath\omega_k(T+\tilde{\ell}-\ell)} = \sum_{k=0}^{T-1} C_k e^{-\imath\omega_{T+\tilde{\ell}-\ell}k} = \tilde{C}_{\tilde{\ell}-\ell}$ for $\tilde{\ell} < \ell$. □

C.6.3 Proof of Lemma C.1

I follow the proof of Theorem 1.3.4 in Ghosh & Ramamoorthi (2003). Set $\kappa_2 = \sup_{\Gamma\in\mathcal{U}^c}\phi(\Gamma)$. Notice that $\hat{\phi}(\Gamma_0) = 0$ and assumption (i) together imply $\phi(\Gamma_0) = 0$. By assumptions (ii)–(iii), we can therefore find a small neighborhood $\mathcal{V}$ of Γ_0 in Ξ_Γ such that $\kappa_1 = \inf_{\Gamma\in\mathcal{V}}\phi(\Gamma)$

satisfies $\max\{\kappa_2, \zeta\} < \kappa_1 < 0$. We may shrink $\mathcal{V}$ to ensure that it also satisfies $\mathcal{V} \subset \mathcal{U} \cap \mathcal{K}$. Choose $\delta > 0$ such that $\kappa_1 - \delta > \max\{\kappa_2 + \delta, \zeta\}$. Write

$$\begin{aligned} P_{\Gamma|Y}(\mathcal{U} \mid Y_T) &= \left(1 + \frac{\int_{\mathcal{U}^c} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}{\int_{\mathcal{U}} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}\right)^{-1} \\ &\geq \left(1 + \frac{\int_{\mathcal{K}^c} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma) + \int_{\mathcal{U}^c \cap \mathcal{K}} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}{\int_{\mathcal{V}} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}\right)^{-1}. \end{aligned}$$

Assumptions (i) and (iv) imply that the following three inequalities hold w.p.a. 1:

$$\inf_{\Gamma \in \mathcal{V}} \hat{\phi}(\Gamma) > \kappa_1 - \delta, \quad \sup_{\Gamma \in \mathcal{U}^c \cap \mathcal{K}} \hat{\phi}(\Gamma) < \kappa_2 + \delta, \quad \sup_{\Gamma \in \mathcal{K}^c} \hat{\phi}(\Gamma) < \zeta.$$

We then have

$$\begin{aligned} P_{\Gamma|Y}(\mathcal{U} \mid Y_T) &\geq \left(1 + \frac{\int_{\mathcal{K}^c} e^{\zeta T} \Pi_\Gamma(d\Gamma) + \int_{\mathcal{U}^c \cap \mathcal{K}} e^{(\kappa_2 + \delta)T} \Pi_\Gamma(d\Gamma)}{\int_{\mathcal{V}} e^{(\kappa_1 - \delta)T} \Pi_\Gamma(d\Gamma)}\right)^{-1} \\ &\geq \left(1 + \frac{e^{\zeta T} + e^{(\kappa_2 + \delta)T}}{\Pi_\Gamma(\mathcal{V}) e^{(\kappa_1 - \delta)T}}\right)^{-1} \end{aligned}$$

w.p.a. 1. Since $\Pi_\Gamma(\mathcal{V}) > 0$ by assumption (v), and $\kappa_1 - \delta > \max\{\kappa_2 + \delta, \zeta\}$, I conclude that $P_{\Gamma|Y}(\mathcal{U} \mid Y_T) \xrightarrow{p} 1$ as $T \to \infty$. □

C.6.4 Proof of Lemma 3

The proof closely follows the steps in Dunsmuir & Hannan (1976, Sec. 3) for proving consistency of the Whittle maximum likelihood estimator in a reduced-form identified VARMA model. Note that the only properties of the data generating process used in Dunsmuir & Hannan (1976, Sec. 3) are covariance stationarity and ergodicity for second moments, as in Assumption 3. Dunsmuir & Hannan also need $T^{-1} y_t y'_{t+T-k} \xrightarrow{p} 0$ for fixed t and k, which follows from Markov's inequality under covariance stationarity. Where Dunsmuir & Hannan (1976) appeal to almost sure convergence, I substitute convergence in probability.

Define the normalized log likelihood ratio

$$\hat{\phi}(\beta, \Sigma) = T^{-1} \log \frac{p^W_{\beta,\Sigma}(Y_T \mid \beta, \Sigma)}{p^W_{\beta,\Sigma}(Y_T \mid \beta_0, \Sigma_0)}.$$

By the Kolmogorov-Szegö formula, for any $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \det(\tilde{f}(\omega; \beta, \Sigma))\, d\omega = \log \det(\Sigma) - n \log(2\pi). \tag{C.2}$$

Hence,

$$\hat{\phi}(\beta, \Sigma) = \frac{1}{2} \log \det(\Sigma_0 \Sigma^{-1}) + \frac{1}{4\pi} \int_{-\pi}^{\pi} \operatorname{tr} \left\{ [\tilde{f}(\omega; \beta_0, \Sigma_0)^{-1} - \tilde{f}(\omega; \beta, \Sigma)^{-1}] \hat{I}(\omega) \right\} d\omega. \tag{C.3}$$

Define also the function

$$\phi(\beta, \Sigma) = \frac{1}{2} \log \det(\Sigma_0 \Sigma^{-1}) + \frac{1}{4\pi} \int_{-\pi}^{\pi} \operatorname{tr}\{I_n - \tilde{f}(\omega; \beta, \Sigma)^{-1} \tilde{f}(\omega; \beta_0, \Sigma_0)\}\, d\omega.$$

$\phi(\beta, \Sigma)$ is continuous. By the argument in Dunsmuir & Hannan (1976, p. 342) (see also Brockwell & Davis, 1991, Prop. 10.8.1, for the univariate case), we have $\phi(\beta, \Sigma) \leq \phi(\beta_0, \Sigma_0) = 0$ for all $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$, with equality if and only if $(\beta, \Sigma) = (\beta_0, \Sigma_0)$.

The remainder of the proof verifies the conditions of Lemma C.1 in four steps.

Step 1. I first show that there exists a neighborhood $\mathcal{K}$ of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$ such that

$$\sup_{(\beta,\Sigma) \in \mathcal{K}} |\hat{\phi}(\beta, \Sigma) - \phi(\beta, \Sigma)| = o_p(1). \tag{C.4}$$

By definition of the Wold decomposition of a time series with a non-singular spectral density, all the roots of $z \mapsto \det(\Phi(\beta_0; z))$ lie strictly outside the unit circle. $\tilde{f}(\omega; \beta, \Sigma)^{-1} = \Phi(\beta; e^{\imath\omega})^{-1\prime} \Sigma^{-1} \Phi(\beta; e^{-\imath\omega})^{-1}$ is therefore uniformly continuous in (ω, β, Σ) for all $\omega \in [-\pi, \pi]$ and (β, Σ) in a small neighborhood of (β_0, Σ_0). Denoting this neighborhood by $\mathcal{K}$, the discussion around Lemma 1 in Dunsmuir & Hannan (1976, p. 350) implies (C.4).

Step 2. For any $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$ and $z \in \mathbb{C}$, define the adjoint of $\Phi(\beta; z)$ as

$$\Phi_{\text{adj}}(\beta; z) = \Phi(\beta; z)^{-1} \det(\Phi(\beta; z)),$$

so $\tilde{f}(\omega; \beta, \Sigma) = |\det(\Phi(\beta; e^{-\imath\omega}))|^2 \Phi_{\text{adj}}(\beta; e^{-\imath\omega})^{-1} \Sigma \Phi_{\text{adj}}(\beta; e^{-\imath\omega})^{-1*}$. The elements of $\Phi_{\text{adj}}(\beta; z)$ are polynomials in z, each polynomial of order at most $\kappa \leq q(n-1)$ (Dunsmuir & Hannan, 1976, p. 354). Write the matrix polynomial as $\Phi_{\text{adj}}(\beta; z) = I_n + \sum_{\ell=1}^{\kappa} \beta_{\text{adj},\ell} z^{\ell}$, and define $\tilde{\Phi}_{\text{adj}}(\beta) = (\sum_{\ell=1}^{\kappa} \|\beta_{\text{adj},\ell}\|^2)^{1/2}$.

Now define, for $\delta \geq 0$,

$$\tilde{f}_\delta(\omega;\beta,\Sigma) = (|\det(\Phi(\beta;e^{-\imath\omega}))|^2 + \delta)\Phi_{\text{adj}}(\beta;e^{-\imath\omega})^{-1}\Sigma\Phi_{\text{adj}}(\beta;e^{-\imath\omega})^{-1*},$$

$$\hat{\phi}_\delta(\beta,\Sigma) = \frac{1}{2}\log\det(\Sigma_0\Sigma^{-1}) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \text{tr}\left\{[\tilde{f}(\omega;\beta_0,\Sigma_0)^{-1} - \tilde{f}_\delta(\omega;\beta,\Sigma)^{-1}]\hat{I}(\omega)\right\} d\omega,$$

and

$$\phi_\delta(\beta,\Sigma) = \frac{1}{2}\log\det(\Sigma_0\Sigma^{-1}) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \text{tr}\{I_n - \tilde{f}_\delta(\omega;\beta,\Sigma)^{-1}\tilde{f}(\omega;\beta_0,\Sigma_0)\}\, d\omega.$$

Because $\hat{I}(\omega)$ is positive semidefinite for each $\omega \in [-\pi,\pi]$, we have $\hat{\phi}(\beta,\Sigma) \leq \hat{\phi}_\delta(\beta,\Sigma)$ for all $(\beta,\Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$ and $\delta > 0$.

Finally, for any $c_1, c_2, c_3 > 0$, define the set

$$\tilde{\mathcal{K}}(c_1,c_2,c_3) = \{(\beta,\Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n \colon \lambda_{\min}(\Sigma) \geq c_1, \|\Sigma\| \leq c_2, \tilde{\Phi}_{\text{adj}}(\beta) \leq c_3\},$$

where $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ.

The discussion surrounding Lemma 1 in Dunsmuir & Hannan (1976, p. 350) then gives

$$\sup_{(\beta,\Sigma)\in\tilde{\mathcal{K}}(c_1,c_2,c_3)} |\hat{\phi}_\delta(\beta,\Sigma) - \phi_\delta(\beta,\Sigma)| = o_p(1),$$

for any $c_1, c_2, c_3 > 0$. Because $\phi_\delta(\beta,\Sigma)$ is continuous in (β,Σ,δ) at $(\beta = \beta_0, \Sigma = \Sigma_0, \delta = 0)$, and $\phi(\beta,\Sigma) = \phi_{\delta=0}(\beta,\Sigma)$ is uniquely maximized at (β_0,Σ_0),

$$\inf_{\delta>0} \sup_{(\beta,\Sigma)\notin\mathcal{K}} \phi_\delta(\beta,\Sigma) < \phi(\beta_0,\Sigma_0) = 0.$$

I conclude that there exist $\delta > 0$ and $\zeta < 0$ such that, for all $c_1, c_2, c_3 > 0$,

$$\sup_{(\beta,\Sigma)\in\tilde{\mathcal{K}}(c_1,c_2,c_3)\cap\mathcal{K}^c} \hat{\phi}(\beta,\Sigma) \leq \sup_{(\beta,\Sigma)\in\tilde{\mathcal{K}}(c_1,c_2,c_3)\cap\mathcal{K}^c} \hat{\phi}_\delta(\beta,\Sigma) \leq \zeta + o_p(1).$$

Step 3. Let ζ be the scalar found in the previous step. The proof of Theorem 4(i) in Dunsmuir & Hannan (1976, pp. 354–355) (see also the beginning of the proof of their Theorem 3, pp. 352–353) shows that there exist $c_1, c_2, c_3 > 0$ such that

$$\sup_{(\beta,\Sigma)\in(\mathbb{B}_{n,q}\times\mathbb{S}_n)\cap\tilde{K}(c_1,c_2,c_3)^c} \hat{\phi}(\beta,\Sigma) \leq \zeta.$$

STEP 4. Steps 1–3 imply that the sufficient conditions in Lemma C.1 hold. I conclude that $P^W_{\beta,\Sigma|Y}(\tilde{\mathcal{U}} \mid Y_T) \xrightarrow{p} 1$ for any neighborhood $\tilde{\mathcal{U}}$ of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$.

REMARK. Finally, I prove an assertion in Appendix A.4.2: Lemma 3 holds for the *discretized* Whittle likelihood that replaces integrals $(2\pi)^{-1} \int_{-\pi}^{\pi} g(\omega)\, d\omega$ in the definition of $\log p^W_{Y|\beta,\Sigma}(Y_T \mid \beta, \Sigma)$ with sums $T^{-1} \sum_{k=0}^{T-1} g(\omega_k)$, $\omega_k = 2\pi k/T$.

The proof of Theorem 4(ii) of Dunsmuir & Hannan (1976, p. 356) shows that steps 1–3 above carry through if the integral in expression (C.3) is replaced with a discretized sum. The only other effect of discretizing the integrals in the Whittle likelihood is that the discretized version of the Kolmogorov-Szegö formula (C.2) does not hold exactly. Instead,

$$T^{-1} \sum_{j=0}^{T-1} \log\det(\tilde{f}(\omega_j; \beta, \Sigma)) = \log\det(\Sigma) - n\log(2\pi) + T^{-1} \sum_{j=0}^{T-1} \log|\det(\Phi(\beta; e^{-\imath\omega_j}))|^2.$$

The posterior consistency result for the discretized Whittle posterior follows from steps 1–4 above if I show

$$\sum_{j=0}^{T-1} \log|\det(\Phi(\beta; e^{-\imath\omega_j}))|^2 \leq 2nq\log 2 \tag{C.5}$$

for all $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$, and furthermore,

$$\sum_{j=0}^{T-1} \log|\det(\Phi(\beta; e^{-\imath\omega_j}))|^2 = o_p(1) \tag{C.6}$$

uniformly in a small neighborhood of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$.

For any $\beta \in \mathbb{B}_{n,q}$ and $z \in \mathbb{C}$, write $\det(\Phi(z; \beta)) = \det(I_n + \sum_{\ell=1}^{q} \beta_\ell z^\ell) = \prod_{b=1}^{nq}(1 - a_b(\beta)z)$ for some complex scalars $\{a_b(\beta)\}_{1 \leq b \leq nq}$ that depend on β and satisfy $|a_b(\beta)| < 1$ (Brockwell & Davis, 1991, p. 191). From the Taylor series $\log(1-z) = -\sum_{s=1}^{\infty} z^s/s$ (valid for $z \in \mathbb{C}$ inside the unit circle) we get, for all $\beta \in \mathbb{B}_{n,q}$,

$$\sum_{k=0}^{T-1} \log\det(\Phi(e^{-\imath\omega_k}; \beta)) = -\sum_{k=0}^{T-1} \sum_{b=1}^{nq} \sum_{s=1}^{\infty} \frac{(a_b(\beta)e^{-\imath\omega_k})^s}{s} = -\sum_{b=1}^{nq} \sum_{s=1}^{\infty} \frac{(a_b(\beta))^s}{s} \sum_{k=0}^{T-1} e^{-\imath\omega_k s}.$$

Since $\sum_{k=0}^{T-1} e^{-\imath\omega_k s}$ equals T when s is an integer multiple of T, and equals 0 otherwise,

$$\sum_{k=0}^{T-1} \log\det(\Phi(e^{-\imath\omega_k}; \beta)) = -\sum_{b=1}^{nq} \sum_{s=1}^{\infty} \frac{(a_b(\beta))^{sT}}{s} = \sum_{b=1}^{nq} \log\left(1 - (a_b(\beta))^T\right).$$

Hence,

$$\begin{aligned}\sum_{k=0}^{T-1} \log |\det(\Phi(e^{-\imath\omega_k};\beta))|^2 &= \sum_{k=0}^{T-1} \log\det(\Phi(e^{-\imath\omega_k};\beta)) + \sum_{k=0}^{T-1} \log\det(\Phi(e^{-\imath\omega_k};\beta)^*) \\ &= \sum_{b=1}^{nq} \log\left|1-(a_b(\beta))^T\right|^2 \\ &\leq nq\log 4,\end{aligned}$$

where the inequality uses $|1-(a_b(\beta))^T| < 2$. Claim (C.5) follows. For β in a small neighborhood of β_0, $\max_b |a_b(\beta)|$ is uniformly bounded away from 1. This implies claim (C.6). □

References

Beaudry, P. & Portier, F. (2014). News-Driven Business Cycles: Insights and Challenges. *Journal of Economic Literature* 52(4), 993–1074.

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods* (2nd ed.). Springer Series in Statistics. Springer.

Dunsmuir, W. & Hannan, E. J. (1976). Vector Linear Time Series Models. *Advances in Applied Probability* 8(2), 339–364.

Durbin, J. & Koopman, S. (2012). *Time Series Analysis by State Space Methods* (2nd ed.). Oxford Statistical Science Series. Oxford University Press.

Fernald, J. (2014). A Quarterly, Utilization-Adjusted Series on Total Factor Productivity. Federal Reserve Bank of San Francisco Working Paper No. 2012-19.

Forni, M., Gambetti, L. & Sala, L. (2014). No News in Business Cycles. *Economic Journal* 124(581), 1168–1191.

Geweke, J. (2010). *Complete and Incomplete Econometric Models*. The Econometric and Tinbergen Institutes Lectures. Princeton University Press.

Ghosh, J. K. & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer.

Hannan, E. (1970). *Multiple Time Series*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15, 1593–1623.

Lippi, M. & Reichlin, L. (1994). VAR analysis, nonfundamental representations, Blaschke matrices. *Journal of Econometrics* 63(1), 307–325.

Lopes, H. F. & Tobias, J. L. (2011). Confronting Prior Convictions: On Issues of Prior Sensitivity and Likelihood Robustness in Bayesian Analysis. *Annual Review of Economics* 3(1), 107–131.

Magnus, J. & Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3rd ed.). Wiley Series in Probability and Statistics. John Wiley & Sons.

Müller, U. K. (2012). Measuring prior sensitivity and prior informativeness in large Bayesian models. *Journal of Monetary Economics* 59(6), 581–597.

Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, Ch. 5, 113–162. CRC Press.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics 3*, 395–402. Oxford University Press.

Sims, E. R. (2012). News, Non-Invertibility, and Structural VARs. In N. Balke, F. Canova, F. Milani, & M. A. Wynne (Eds.), *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments*, Vol. 28 of *Advances in Econometrics*, 81–135. Emerald Group Publishing.

Stan Development Team (2015). *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*.

Stock, J. H. & Watson, M. W. (2012). Disentangling the Channels of the 2007–09 Recession. *Brookings Papers on Economic Activity* (Spring), 81–135.