Prior Selection for Vector Autoregressions

Domenico Giannone
Université Libre de Bruxelles and CEPR

Michele Lenza
European Central Bank

Giorgio E. Primiceri
Northwestern University, CEPR and NBER

ECARES working paper 2012-002

ECARES
ULB - CP 114
50, F.D. Roosevelt Ave., B-1050 Brussels BELGIUM
www.ecares.org

Prior Selection for Vector Autoregressions*

Domenico Giannone (Université Libre de Bruxelles and CEPR)
Michele Lenza (European Central Bank)
Giorgio E. Primiceri (Northwestern University, CEPR and NBER)

First Version: March 2010
This Version: December 2011

Abstract

Vector autoregressions (VARs) are flexible time series models that can capture complex dynamic interrelationships among macroeconomic variables. However, their dense parameterization leads to unstable inference and inaccurate out-of-sample forecasts, particularly for models with many variables. A potential solution to this problem is to use informative priors, in order to shrink the richly parameterized unrestricted model towards a parsimonious naïve benchmark, and thus reduce estimation uncertainty. This paper studies the optimal choice of the informativeness of these priors, which we treat as additional parameters, in the spirit of hierarchical modeling. This approach is theoretically grounded, easy to implement, and greatly reduces the number and importance of subjective choices in the setting of the prior. Moreover, it performs very well both in terms of out-of-sample forecasting, and accuracy in the estimation of impulse response functions.

1 Introduction

In this paper, we study the choice of the informativeness of the prior distribution on the coefficients of the following VAR model:

$$\begin{aligned} y_t &= C + B_1 y_{t-1} + ... + B_p y_{t-p} + \varepsilon_t \\ \varepsilon_t &\sim N(0, \Sigma), \end{aligned} \tag{1.1}$$

where y_t is an $n \times 1$ vector of endogenous variables, ε_t is an $n \times 1$ vector of exogenous shocks, and C, B_1,..., B_p and Σ are matrices of suitable dimensions containing the model's unknown parameters.

With flat priors and conditioning on the initial p observations, the posterior distribution of $\beta \equiv vec\left([C, B_1, ..., B_p]'\right)$ is centered at the Ordinary Least Square (OLS) estimate of the coefficients and it is easy to compute. It is well known, however,

*We thank Liseo Brunero, Guenter Coenen, Gernot Doppelhofer, Raffaella Giacomini, Dimitris Korobilis, Frank Schorfheide, Chris Sims and participants in several conferences and seminars for comments and suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Eurosystem.

that working with flat priors leads to inadmissible estimators (Stein, 1956) and yields poor inference, particularly in large dimensional systems (see, for example, Sims, 1980; Litterman, 1986; Bańbura, Giannone, and Reichlin, 2010; Koop and Korobilis, 2010). One typical symptom of this problem is the fact that these models generate inaccurate out-of-sample predictions, due to the large estimation uncertainty of the parameters.

To improve the forecasting performance of VAR models, the literature has proposed to combine the likelihood function with some informative prior distributions. Using the frequentist terminology, these priors are successful because they effectively reduce the estimation error, while generating only relatively small biases in the estimates of the parameters. To illustrate this point more formally from a Bayesian perspective, let's consider the following (conditional) prior distribution for the VAR coefficients

$$\beta|\Sigma \sim N\left(b, \Sigma \otimes \Omega\xi\right),$$

where b and Ω are given, and ξ is a scalar controlling the tightness of the prior information. The conditional posterior of β can be obtained by multiplying this prior by the likelihood function, and takes the form

$$\begin{aligned}
\beta|\Sigma, y &\sim N\left(\hat{\beta}\left(\xi\right), \hat{V}\left(\xi\right)\right) \\
\hat{\beta}\left(\xi\right) &\equiv vec\left(\hat{B}\left(\xi\right)\right) \\
\hat{B}\left(\xi\right) &\equiv \left(x'x + \left(\Omega\xi\right)^{-1}\right)^{-1}\left(x'y + \left(\Omega\xi\right)^{-1}b\right) \\
\hat{V}\left(\xi\right) &\equiv \Sigma \otimes \left(x'x + \left(\Omega\xi\right)^{-1}\right)^{-1},
\end{aligned}$$

where $y \equiv [y_{p+1}, ..., y_T]'$ is the $(T-p) \times n$ matrix of observed data up to time T, $x \equiv [x_{p+1}, ..., x_T]'$ and $x_t \equiv [1, y'_{t-1}, ..., y'_{t-p}]'$. Notice that, if we choose a lower ξ, the prior becomes more informative, the posterior mean of β comes closer to the prior mean, and the posterior variance falls.

One natural way to assess the impact of different priors on the model's ability to fit the data is to evaluate their effect on the model's out-of-sample forecasting performance, summarized by the probability of observing low forecast errors. To this end, rewrite (1.1) as

$$y_t = X_t\beta + \varepsilon_t,$$

where $X_t \equiv I_n \otimes x'_t$ and I_n denotes an $n \times n$ identity matrix. The distribution of the one-step-ahead forecast is then given by

$$y_{T+1}|\Sigma, y \sim N\left(X_T\hat{\beta}\left(\xi\right),\ X_T\hat{V}\left(\xi\right)X'_T + \Sigma\right),$$

whose variance depends both on the posterior variance of the coefficients and the volatility of the innovations. It is then easy to see that neither very high nor very low values of ξ are likely to be ideal. On the one hand, if ξ is too low and the prior very dogmatic, density forecasts will be very concentrated around $X_T b$. This results in a low probability of observing small forecast errors, unless the prior mean happens to be in a close neighborhood of the likelihood peak (and there is no reason to believe that this is the

case, in general). On the other hand, if ξ is too high and the prior too uninformative, the model generates very dispersed density forecasts, especially in high-dimensional VARs, because of high estimation uncertainty. This also lowers the probability of observing small forecast errors, despite the fact that the distance between y_{T+1} and $X_t\hat{\beta}$ might be small. In sum, neither flat nor dogmatic priors maximize the fit of the model, which makes the choice of the informativeness of the prior distribution a crucial issue.

The literature has proposed a number of heuristic methodologies to set the informativeness of the prior distribution. In the context of VARs, for example, Litterman (1980) and Doan, Litterman, and Sims (1984) set the tightness of the prior by maximizing the out-of-sample forecasting performance of the model. Bańbura, Giannone, and Reichlin (2010) propose instead to control for over-fitting by choosing the shrinkage parameters that yield a desired in-sample fit.[1]

From a purely Bayesian perspective, however, the choice of the informativeness of the prior distribution is conceptually identical to the inference on any other unknown parameter of the model. Suppose, for instance, that a model is described by a likelihood function $p(y|\theta)$ and a prior distribution $p_\gamma(\theta)$, where θ are the model's parameters and γ corresponds to the hyperparameters, i.e. those coefficients that parameterize the prior distribution, but do not directly affect the likelihood.[2] It is then natural to choose these hyperparameters by interpreting the model as a hierarchical model, i.e. replacing $p_\gamma(\theta)$ with $p(\theta|\gamma)$, and evaluating their posterior (Berger, 1985; Koop, 2003). Such a posterior can be obtained by applying Bayes' law, which yields

$$p(\gamma|y) \propto p(y|\gamma) \cdot p(\gamma),$$

where $p(\gamma)$ denotes the prior density on the hyperparameters—also known as the hyperprior—while $p(y|\gamma)$ is the so called marginal likelihood (ML), and corresponds to

$$p(y|\gamma) = \int p(y|\theta,\gamma)\, p(\theta|\gamma)\, d\theta. \tag{1.2}$$

In other words, the ML is the density of the data as a function of the hyperparameters γ, obtained after integrating out the uncertainty about the model's parameters θ. Conveniently, in the case of VARs with conjugate priors, the ML is also available in closed form.

Notice, also, that the hierarchical prior structure implies that the unconditional prior for the parameters θ has a mixed distribution

$$p(\theta) = \int p(\theta|\gamma)\, p(\gamma)\, d\gamma.$$

Mixed distributions have generally fatter tails than each of the component distributions $p(\theta|\gamma)$, a property that robustifies inference. In fact, when the prior has fatter tails

[1] A number of papers have subsequently followed either the first (e.g. Robertson and Tallman, 1999; Wright, 2009; Giannone, Lenza, Momferatou, and Onorante, 2010) or the second strategy (e.g. Giannone, Lenza, and Reichlin, 2008; Bloor and Matheson, 2009; Carriero, Kapetanios, and Marcellino, 2009; Koop, 2011).

[2] The distinction between parameters and hyperparameters is mostly fictitious and made only for convenience.

than the likelihood, the posterior is less sensitive to extreme discrepancies between prior and likelihood (Berger, 1985; Berger and Berliner, 1986).

Conducting formal inference on the hyperparameters is theoretically grounded and has also several appealing interpretations. For example, with a flat hyperprior, the shape of the posterior of the hyperparameters coincides with the ML, which is a measure of out-of-sample forecasting performance of a model (see Geweke, 2001; Geweke and Whiteman, 2006). More specifically, the ML corresponds to the probability density that the model generates zero forecast errors, which can be seen by rewriting the ML as a product of conditional densities:

$$p\left(y|\gamma\right) = p\left(y_1|\gamma\right) \cdot \prod_{t=2}^{T} p\left(y_t|y^{t-1},\gamma\right).$$

As a consequence, maximizing the posterior of the hyperparameters corresponds to maximizing the one-step-ahead out-of-sample forecasting ability of the model.

Moreover, the strategy of estimating hyperparameters by maximizing the ML (i.e. their posterior under a flat hyperprior) is an Empirical Bayes method (Robbins, 1956), which has a clear frequentist interpretation. On the other hand, the full posterior evaluation of the hyperparameters (as advocated, for example, by Lopes, Moreira, and Schmidt, 1999, for VARs) can be thought of as conducting Bayesian inference on the population parameters of a random effects model or, more generally, of a hierarchical model (see, for instance, Gelman, Carlin, Stern, and Rubin, 2004).

In this paper, we adopt this hierarchical modeling approach to make inference about the informativeness of the prior distribution of Bayesian Vector Autoregressions (BVARs) estimated on postwar U.S. macroeconomic data. We consider a combination of the conjugate priors most commonly used in the literature (the "Minnesota," "sum of coefficients" and "dummy initial observation" priors), and document that this estimation strategy generates very accurate out-of-sample predictions, both in terms of point and density forecasts. The key to success lies in the fact that this procedure automatically selects the "appropriate" amount of shrinkage, namely tighter priors when the model involves many unknown coefficients relative to the available data, and looser priors in the opposite case. Because of this feature, the hierarchical BVAR improves over naive benchmarks and flat-prior VARs, even for small-scale models, for which the optimal shrinkage is low, but not zero. In addition, we find that the forecasting performance of the model normally improves as we include more variables, and it is comparable to factor models, which are among the most successful methods to deal with large sets of predictors.

Our second contribution is documenting that this hierarchical BVAR approach performs very well also in terms of accuracy of the estimation of impulse response functions in identified VARs. We conduct two experiments to make this point. First, we study the transmission of an exogenous increase in the federal funds rate in a large-scale model with 22 variables. The estimates of the impulse responses that we obtain are broadly in line with the usual narrative of the effects of an exogenous tightening in monetary policy. This finding, together with the result that the same large-scale model produces good forecasts, indicates that our approach is able to effectively deal with the

curse of dimensionality. However, in this empirical exercise there is no way of formally checking the accuracy of the estimated impulse response functions, since we do not have a directly observable counterpart of these objects in the data. Therefore, we conduct a second exercise, which is a controlled Monte Carlo experiment. Namely, we simulate data from a micro-founded, medium-scale, dynamic stochastic general equilibrium model estimated on U.S. postwar data. We then use the simulated data to estimate our hierarchical BVAR, and compare the implied impulse responses to monetary policy shocks to those of the true data generating process. This experiment lends strong support to our model. The surprising finding is in fact that the hierarchical Bayesian procedure generates very little bias, while drastically increasing the efficiency of the impulse response estimates relative to standard flat-prior VARs.

Hierarchical modeling (or Empirical Bayes, i.e. its frequentist version) has been successfully adopted in many fields (see Berger, 1985; Gelman, Carlin, Stern, and Rubin, 2004, for an overview). It has also been advocated by the first proponents of BVARs (Doan, Litterman, and Sims, 1984; Sims and Zha, 1998; Canova, 2007) but seldom formally implemented in this context.[3] Exceptions to this statement include Del Negro and Schorfheide (2004) and Del Negro, Schorfheide, Smets, and Wouters (2007), who use the ML to choose the tightness of a prior for VARs derived from the posterior density of a dynamic stochastic general equilibrium model. Relative to these authors, our focus is on BVARs with standard conjugate priors, for which the posterior of the hyperparameters is available in closed form. Phillips (1995) chooses the hyperparameters of the Minnesota prior using the asymptotic posterior odds criterion of Phillips and Ploberger (1994), which is also related to the ML. More recently, Carriero, Kapetanios, and Marcellino (2010) and Carriero, Clark, and Marcellino (2011) have used the ML to set the variance of a (variant of the) Minnesota prior for VARs. They show that such a strategy is successful in forecasting bond yields and macroeconomic variables.

We generalize this approach to the optimal selection of a variety of commonly adopted prior distributions for BVARs. This includes the prior on the sum of coefficients proposed by Doan, Litterman, and Sims (1984), which turns out to be crucial to enhance the forecasting performance of the model. Moreover, relative to these studies, we take an explicit hierarchical modeling approach that allows us to take the uncertainty about hyperparameters into account, and to evaluate the density forecasts of the model. More important, we also complement the model's forecasting evaluation with an assessment of the performance of hierarchical BVARs for impulse response estimation, which is new in the literature. Finally, we document that our approach works well for models of very different scale, including 3-variable VARs and much larger-scale ones. In this respect, our work relates to the growing literature on forecasting using factors extracted from large information sets (see, for example Forni, Hallin, Lippi, and Reichlin, 2000; Stock and Watson, 2002b), Large Bayesian VARs (Bańbura, Giannone, and Reichlin, 2010) and empirical Bayes regressions with large sets of predictors (Knox, Stock, and Watson, 2000).

The rest of the paper is organized as follows. Section 2 and 3 provide some addi-

[3] In the context of time varying VARs, hierarchical modeling has been used by Primiceri (2005) and Belmonte, Koop, and Korobilis (2011) to choose the informativeness of the prior distribution for the time variation of coefficients and volatilities.

tional details about the computation and interpretation of the ML, and the priors and hyperpriors used in our investigation. Section 4 and 5 focus instead on the empirical application to macroeconomic forecasting and impulse response estimation. Section 6 concludes.

2 The Choice of Hyperparameters for BVARs

In the previous section, we have argued that the most natural way of choosing the hyperparameters of a model is based on their posterior distribution. This posterior is proportional to the product of the hyperprior and the ML. The hyperprior is a "level-two" prior on the hyperparameters, while the ML is the likelihood of the observed data as a function of the hyperparameters, which we can obtain by integrating out the model's coefficients, as in equation (1.2). Although this procedure can be applied very generally, in this paper we restrict our attention to prior distributions for VAR coefficients belonging to the following Normal-Inverse-Wishart family:

$$\Sigma \sim IW(\Psi; d) \tag{2.3}$$
$$\beta|\Sigma \sim N(b, \Sigma \otimes \Omega), \tag{2.4}$$

where the elements Ψ, d, b and Ω are typically functions of a lower dimensional vector of hyperparameters γ.

We focus on these priors for two reasons. First of all, this class includes the priors most commonly used by the existing literature on BVARs.[4] Second, the prior (2.3)-(2.4) is conjugate and has the advantage that the ML can be computed in closed form as a function of γ. In appendix A we show that the posterior of the parameters has the following Normal-Inverse-Wishart distribution:

$$\Sigma|y \sim IW\left(\Psi + \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B}-\hat{b}\right)'\Omega^{-1}\left(\hat{B}-\hat{b}\right), T-p+d\right) \tag{2.5}$$
$$\beta|\Sigma, y \sim N\left(\hat{\beta}, \Sigma \otimes \left(x'x+\Omega^{-1}\right)^{-1}\right), \tag{2.6}$$

and that the ML is given by the following expression:

$$\begin{aligned} p(y|\gamma) = & \left(\frac{1}{\pi}\right)^{\frac{n(T-p)}{2}} \frac{\Gamma_n\left(\frac{T-p+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)} \cdot \\ & |\Omega|^{-\frac{n}{2}} \cdot |\Psi|^{\frac{d}{2}} \cdot \left|x'x+\Omega^{-1}\right|^{-\frac{n}{2}} \cdot \\ & \left|\Psi + \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B}-\hat{b}\right)'\Omega^{-1}\left(\hat{B}-\hat{b}\right)\right|^{-\frac{T-p+d}{2}}, \end{aligned} \tag{2.7}$$

where $\Gamma_n(\cdot)$ is the n-variate Gamma function, $\hat{\varepsilon}$ is the $(T-p)\times n$ matrix of the VAR residuals computed at the posterior mode of the VAR parameters, $\hat{b}$ is an $(1+np)\times n$

[4] Some recent studies have proposed alternative priors for VAR that do not belong to this family. See, for example, Del Negro and Schorfheide (2004), Villani (2009) and Jarociski and Marcet (2010).

matrix obtained by reshaping the vector b in such a way that each column corresponds to the prior mean of the coefficients of each equation, and $\hat{B}$ is an $(1+np)\times n$ matrix obtained by reshaping the posterior mode of the VAR coefficients in such a way that each column corresponds to the posterior mode of the coefficients of each equation. Notice that (2.7) is a function of γ because in principle Ψ, d, b, Ω, $\hat{\varepsilon}$ and $\hat{B}$ all depend on γ.[5]

Apart from a constant term, the logarithm of the ML can be written as the sum of two components. The first component is the difference between the log-determinant of the prior and posterior mode (or mean) of the residual covariance matrix, weighted by their respective degrees of freedom. The second part consists of the difference between the log-determinant of the prior and posterior variance of the model's coefficients. On the one hand, less informative priors improve the ML to the extent that they increase the in-sample fit of the model and reduce the posterior mode of the residual covariance matrix. On the other hand, weaker prior information typically induces a greater discrepancy between the prior and posterior variance of the coefficients, which penalizes the ML.

Clearly, the fact that the ML is available in closed form simplifies inference substantially. Given (2.7), it is in fact easy to either maximize or simulate the posterior of the hyperparameters. As we have pointed out in the introduction, the advantage of the approach based on the maximization is that, under a flat hyperprior, it is an Empirical Bayes procedure and has a classical interpretation. In addition, it coincides with selecting hyperparameters that maximize the one-step-ahead out-of-sample forecasting performance of the model. On the other hand, the full posterior simulation allows to account for the estimation uncertainty of the hyperparameters, and has an interpretation of Bayesian hierarchical modeling. This approach can be implemented using a simple Markov chain Monte Carlo algorithm. In particular, we use a Metropolis step to draw the low dimensional vector of hyperparameters. Conditional on a value of γ, the VAR coefficients $[\beta, \Sigma]$ can then be drawn from their posterior, which is Normal-Inverse-Wishart and given by (6.10)-(6.11). Appendix B presents the details of this procedure.

We now turn to the description of the specific priors that we employ in our empirical analysis.

3 Priors and Hyperpriors

As mentioned in the previous section, we focus on priors of the form (2.3)-(2.4). As in Kadiyala and Karlsson (1997), we set the degrees of freedom of the Inverse-Wishart distribution to $d = n+2$, which is the minimum value that guarantees the existence of the prior mean of Σ (it is equal to $\Psi/(d-n-1)$). In addition, we take Ψ to be a diagonal matrix with an $n\times 1$ vector ψ on the main diagonal. We treat ψ as an

[5] It is common in the literature to implement some of these conjugate priors using dummy observations. In this case, abusing notation, the ML would be given by $p(y|\gamma) = p\left(y^{+}|\gamma\right) - p\left(y^{*}|\gamma\right)$, where $p(\cdot|\cdot)$ is the function in (2.7), y^{*} denote the artificial data, and y^{+} is the extended set of data, i.e. $[y, y^{*}]$.

hyperparameter, which differs from the existing literature that has been fixing this parameter using sample information.

As for the conditional Gaussian prior for β, we combine the three most popular prior densities used by the existing literature for the estimation of BVARs in levels:

1. The baseline prior is a version of the so-called Minnesota prior, first introduced in Litterman (1979, 1980). This prior is centered on the assumption that each variable follows a random walk process, possibly with drift, which is a parsimonious yet "reasonable approximation of the behavior of an economic variable"(Litterman, 1979, p. 20). More precisely, this prior is characterized by the following first and second moments:

$$\begin{aligned} E\left[(B_s)_{ij}|\Sigma\right] &= \begin{cases} 1 & \text{if } i=j \text{ and } s=1 \\ 0 & \text{otherwise} \end{cases} \\ cov\left((B_s)_{ij},(B_r)_{hm}|\Sigma\right) &= \begin{cases} \lambda^2 \frac{1}{s^2} \frac{\Sigma_{ih}}{\psi_j/(d-n-1)} & \text{if } m=j \text{ and } r=s \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

   and can be easily cast into the form of (2.4). Notice that the variance of this prior is lower for the coefficients associated with more distant lags, and that coefficients associated with the same variable and lag in different equations are allowed to be correlated. Finally, the key hyperparameter is λ, which controls the scale of all the variances and covariances, and effectively determines the overall tightness of this prior.

The literature following Litterman's work has introduced refinements of the Minnesota prior to further "favor unit roots and cointegration, which fits the beliefs reflected in the practices of many applied macroeconomists" (Sims and Zha, 1998, p. 958). Loosely speaking, the objective of these additional priors is to reduce the importance of the deterministic component implied by VARs estimated conditioning on the initial observations (Sims, 1992a). This deterministic component is defined as $\tau_t \equiv E_p\left(y_t|y_1,...,y_p,\hat{\beta}\right)$, i.e. the expectation of future y's given the initial conditions and the value of the estimated VAR coefficients. According to Sims (1992a), in unrestricted VARs, τ_t has a tendency to exhibit temporal heterogeneity—a markedly different behavior at the beginning and the end of the sample—and to explain an implausibly high share of the variation of the variables over the sample. As a consequence, priors limiting the explanatory power of this deterministic component have been shown to improve the forecasting performance of BVARs.

2. The first prior of this type is known as "sum-of-coefficients" prior and was originally proposed by Doan, Litterman, and Sims (1984). Following the literature, it is implemented using Theil mixed estimation, with a set of n artificial observations—one for each variable—stating that a no-change forecast is a good forecast at the beginning of the sample. More precisely, we construct the following

set of dummy observations:

$$\begin{aligned} \underset{n\times n}{y^{+}} &= diag\left(\frac{\bar{y}_0}{\mu}\right) \\ \underset{n\times(1+np)}{x^{+}} &= \left[\underset{n\times 1}{0}, y^{+}, ..., y^{+}\right], \end{aligned}$$

where $\bar{y}_0$ is an $n \times 1$ vector containing the average of the first p observations for each variable, and the expression $diag(v)$ denotes the diagonal matrix with the vector v on the main diagonal. These artificial observations are added on top of the data matrices $y \equiv [y_{p+1}, ..., y_T]'$ and $x \equiv [x_{p+1}, ..., x_T]'$, which are then used for inference. The prior implied by these dummy observations is centered at 1 for the sum of coefficients on own lags for each variable, and at 0 for the sum of coefficients on other variables' lags. It also introduces correlation among the coefficients on each variable in each equation. The hyperparameter μ controls the variance of these prior beliefs: as $\mu \to \infty$ the prior becomes uninformative, while $\mu \to 0$ implies the presence of a unit root in each equation and rules out cointegration.

3. The fact that, in the limit, the sum-of-coefficients prior is not consistent with cointegration motivates the use of an additional prior that was introduced by Sims (1993), known as "dummy-initial-observation" prior. It is implemented using the following dummy observation

$$\begin{aligned} \underset{1\times n}{y^{++}} &= \frac{\bar{y}_0'}{\delta} \\ \underset{1\times(1+np)}{x^{++}} &= \left[\frac{1}{\delta}, y^{++}, ..., y^{++}\right], \end{aligned}$$

which states that a no-change forecast *for all variables* is a good forecast at the beginning of the sample. The hyperparameter δ controls the tightness of the prior implied by this artificial observation. As $\delta \to \infty$ the prior becomes uninformative. On the other hand, as $\delta \to 0$, all the variables of the VAR are forced to be at their unconditional mean, or the system is characterized by the presence of an unspecified number of unit roots without drift. As such, the dummy-initial-observation prior is consistent with cointegration.

Summing up, the setting of these priors depends on the hyperparameters λ, μ, δ and ψ, which we treat as additional parameters. As hyperpriors for λ, μ and δ, we choose Gamma densities with mode equal to 0.2, 1 and 1—the values recommended by Sims and Zha (1998)—and standard deviations equal to 0.4, 1 and 1 respectively. Finally, our prior on $\psi/(d-n-1)$, i.e. the prior mean of the main diagonal of Σ, is an Inverse-Gamma with scale and shape equal to $(0.02)^2$. This prior peaks at approximately $(0.02)^2$, is proper, but quite disperse since it does not have neither a variance nor a mean. We work with proper hyperpriors because they guarantee the properness of the posterior and, from a frequentist perspective, the admissibility of the estimator of the hyperparameters, which is a difficult property to check for the case of hierarchical

models (see Berger, Strawderman, and Dejung, 2005). However, our hyperpriors are relatively diffuse, and our empirical results are confirmed when using completely flat, improper hyperpriors.

4 Forecasting Evaluation of BVAR Models

The assessment of the forecasting performance of econometric models has become standard in macroeconomics, even when the main objective of the study is not to provide accurate out-of-sample predictions. This is because the forecasting evaluation can be thought of as a model validation procedure. In fact, if model complexity is introduced with a proliferation of parameters, instabilities due to estimation uncertainty might completely offset the gains obtained by limiting model mis-specification. Out-of-sample forecasting reflects both parameter uncertainty and model mis-specification and reveals whether the benefits due to flexibility are outweighed by the fact that the more general model captures also non-prominent features of the data.

Our out-of-sample evaluation is based on the US dataset constructed by Stock and Watson (2008). We work with three different VAR models, including progressively larger sets of variables:[6]

1. A *SMALL*-scale model—the prototypical monetary VAR—with three variables, i.e. GDP, the GDP deflator and the federal funds rate.

2. A *MEDIUM*-scale model, which includes the variables used for the estimation of the DSGE model of Smets and Wouters (2007) for the US economy. In other words, we add consumption, investment, hours worked and wages to the small model.

3. A *LARGE*-scale model, with 22 variables, using a dataset that nests the previous two specifications and also includes a number of important additional labor market, financial and monetary variables.

Further details on the database are reported in Table 1.

INSERT TABLE 1 HERE

The variables enter the models in annualized log-levels (i.e. we take logs and multiply by 400), except those already defined in terms of annualized rates, such as interest rates, which are taken in levels. The number of lags in all the VARs is set to five.

Using each of these three datasets, we produce the BVAR forecasts recursively for two horizons (1 and 4 quarters), starting with the estimation sample that ranges from 1959Q1 to 1974Q4. More precisely, using data from 1959Q1 to 1974Q4, we generate draws from the posterior predictive density of the model for 1975Q1 (one quarter ahead) and 1975Q4 (one year ahead). We then iterate the same procedure updating

[6]The complete database in Stock and Watson (2008) includes 149 quarterly variables from 1959Q1 to 2008Q4. Since several variables are monthly, we follow Stock and Watson (2008) and transform them into quarterly by taking averages.

the estimation sample, one quarter at a time, until the end of the sample, i.e. 2008Q4. At each iteration, of course, we also re-estimate the posterior distribution of the hyperparameters. The outcome of this procedure is a time-series of 137 density forecasts for each of the two forecast horizons.

We start by assessing the accuracy of our models in terms of point forecasts, defined as the median of the predictive density at each point in time. We then turn to the evaluation of the density forecasts to assess how accurately different models capture the uncertainty around the point forecasts.

For each variable, the target of our evaluation is defined in terms of the h-period annualized average growth rates, i.e. $z_{i,t+h}^{h} = \frac{1}{h}[y_{i,t+h} - y_{i,t}]$. For variables specified in log-levels, this is approximately the average annualized growth rate over the next h quarters, while for variables not transformed in logs this is the average quarterly change over the next h quarters.

We compare the forecasting performance of the BVAR to a VAR with flat prior, estimated by OLS (we will refer to this model as VAR or flat-prior VAR) and a random walk with drift, which is the model implied by a dogmatic Minnesota prior (we will refer to this model as RW). We also compare the point forecasts of the BVAR to those of a single equation model, augmented with factors extracted from a large dataset using principal components.[7] Factor models offer a parsimonious representation for macroeconomic variables while retaining the salient features of the data that notoriously strongly comove. Hence, factor augmented regressions are widely used in order to deal with the curse of dimensionality, since a large set of potential predictors can be replaced in the regressions by a much smaller number of factors. Factor based approaches are a benchmark in the literature and have been shown to produce very accurate forecasts exploiting large cross-sections of data. Specifically we focus on the factor based forecasting approach of Stock and Watson (2002a,b), whose implementation details are reported in appendix C.

4.1 Point forecasts

Table 1 analyzes the accuracy of point forecasts by reporting the mean squared forecast errors (MSFE) of real GDP, the GDP deflator and the federal funds rate.

INSERT TABLE 2 HERE

Comparing models of different size, notice that it is not possible to estimate the large-scale VAR with a flat prior. In addition, the VAR forecasts worsen substantially when moving from the small to the medium-scale model. This outcome indicates that the gains from exploiting larger information sets are completely offset by an increase in estimation error. On the contrary, the forecast accuracy of the BVARs does not deteriorate when increasing the scale of the model, and sometimes even improves substantially (as it is the case for inflation). In this sense, the use of priors seems to be able to turn the curse into a blessing of dimensionality. Moreover, BVAR forecasts are

[7] The principal components are extracted from the whole set of 149 variables described in Stock and Watson (2008).

systematically more accurate than the flat-prior VAR forecasts, for all the variables and horizons that we consider.

The comparison with the RW model is also favorable to the BVARs, with the possible exception of the forecasts of the federal funds rate at one-year horizon. The improvement of BVARs over the prior model indicates that our inference-based choice of the hyperparameters leads to the use of informative priors, but not excessively so, letting the data shape the posterior beliefs about the model's coefficients. Finally, notice that the performance of the prior model is particularly poor for inflation. In fact Atkeson and Ohanian (2001) show that a random walk for the *growth rate* of the GDP deflator is a more appropriate naive benchmark model. Specifically, they propose to forecast inflation over the subsequent year using the inflation rate over the past year. The MSFE of this alternative simple model for inflation at a 4-quarter horizon is 1.24, which is smaller than that obtained with the random walk in levels or with the small and medium BVARs, but higher than the corresponding MSFE of the large-scale BVAR.

Table 2 also suggests that the BVAR predictions are competitive with those of the factor model. This outcome is in line with the findings of De Mol, Giannone, and Reichlin (2008) and indicates that factor augmented and Bayesian regressions capture the same features of the data. In fact, De Mol, Giannone, and Reichlin (2008) have shown that Bayesian shrinkage and regressions augmented with principal components are strictly connected.

Overall, the results presented in table 2 are in line with the conclusion of existing studies that highlight the accuracy of BVAR forecasts (Doan, Litterman, and Sims, 1984; Sims and Zha, 1998; Robertson and Tallman, 2001; Bańbura, Giannone, and Reichlin, 2010; Koop, 2011), although these authors select the tightness of the prior information using heuristic procedures. This suggests that BVARs improve forecast accuracy over models with flat priors across a relatively wide range of parameter settings. However, "the degree of forecast accuracy improvement for a given data set is dependent on the choice of hyperparameter values" (Robertson and Tallman, 2001, p.14), an issue that gives support to our inference-based procedure for choosing the hyperparameters. Another advantage of our methodology relative to more ad hoc procedures is that it can be used with different sets of data, without requiring human judgement in the search for reasonable ranges of hyperparameters.

4.2 Density forecasts

The point forecast evaluation of the previous subsection is a useful tool to discriminate among models, but disregards the uncertainty assigned by each model to its point prediction. For this reason, we now turn to the evaluation of density forecasts. We measure the accuracy of a density forecast using the log-predictive score, which is simply the logarithm of the predictive density generated by a model, evaluated at the realized value of the time series. Therefore, if model A has a higher average log predictive score than model B, it means that values close to the actual realizations of a time series were a priori more likely according to model A relative to model B.

Table 3 reports the average difference between the log predictive scores of the

BVARs and the competing models (the flat-prior VAR and RW models), for each variable and horizon. A positive number indicates that the density forecasts produced by our proposed procedure are superior to those of the alternative models. In addition, the HAC estimate of its standard deviation (in parentheses) gives a rough idea of the statistical significance and the volatility of this difference.[8]

INSERT TABLE 3 HERE

Table 3 makes clear that the BVAR forecasts are more accurate than those of the RW and flat-prior VAR also when evaluating the whole density.

4.3 Inspecting the mechanism

In this subsection, we provide some intuition about why the hierarchical procedure described in the previous sections generates accurate forecasts. As we have discussed at length in the introduction, VAR models require the estimation of many free parameters, which, when using a flat prior, leads to high estimation uncertainty and overfitting. It is therefore beneficial to shrink the model parameters towards a parsimoneous prior model. The key to success of the hierarchical BVAR is that it automatically infers the "appropriate" amount of shrinkage, by selecting the tightness of the prior distribution. For example, the procedure will select looser priors for models with fewer parameters, and tighter priors for models with many parameters relative to the available data.

To illustrate this point, consider a much simplified version of our model, i.e a BVAR with only a Minnesota prior, and the prior mean of the diagonal elements of Σ set equal to the variance of the residuals of an AR(1) for each variables (as in Kadiyala and Karlsson, 1997). This model is convenient because it involves only one hyperparameter, namely the hyperparameter λ governing the overall standard deviation of the Minnesotra prior. For each dataset—small, medium and large—we estimate our hierarchical BVAR on the full sample, and compute the posterior distribution of the hyperparameter λ. These posteriors are plotted in figure 1, along with the hyperprior. Notice that, in line with intuition, the posterior mode (and variance) of λ decreases with the size of the model. In other words, the larger the size of the BVAR, the more likely it is that we should shrink the model toward the parsimonious specification implied by the Minnesota prior.

INSERT FIGURE 1 HERE

5 Structural BVARs and estimation of impulse response functions

The forecast accuracy of the hierarchical modeling procedure proposed in this paper is quite remarkable, and in line with the interpretation of the marginal likelihood as a

[8] Notice that the associated t-statistics corresponds to the statistics of Amisano and Giacomini (2007) with standard Normal distribution when the models are estimated using a rolling scheme. This is not the case in our exercise since we use a recursive estimation procedure.

measure of out-of-sample forecasting performance. However, VARs are not used in the literature only for forecasting, but also as a tool to identify structural shocks and assess their transmission mechanism. Inspired by an important insight of statistical decision theory—the separation between loss functions and probability models—we now present evidence that the same hierarchical modeling strategy also delivers accurate estimates of the impulse response functions to structural shocks.

More specifically, in this section we perform two exercises. First, we estimate the impulse responses to monetary policy shocks using our large-scale BVAR with 22 variables. The analysis of the effects of monetary policy innovations is widespread in the literature because, among other things, it allows to discriminate between competing theoretical models of the economy (Christiano, Eichenbaum, and Evans, 1999). The purpose of this first exercise is to demonstrate that our hierarchical procedure allows us to obtain plausible estimates of impulse response functions even when working with large-scale models, which is not the case for flat-prior VARs. However, we do not have an observable counterpart of these impulse responses in the data that can be used to directly check their accuracy. This motivates our second exercise, which is a controlled Monte Carlo experiment. In a nutshell, we simulate artificial datasets from a dynamic stochastic general equilibrium (DSGE) model, and assess the gains in accuracy for the estimation of impulse responses to monetary policy shocks of our hierarchical procedure over flat-prior VARs.

Concerning our first exercise, the monetary policy shock is identified using a relatively standard recursive identification scheme, assuming that prices and real activity do not react contemporaneously to the monetary policy shock. The only variables that can react contemporaneously to monetary policy shocks are the financial variables (bond rates and stock prices), the exchange rate and M2, while the policy rate does not react contemporaneously to financial variables (see Christiano, Eichenbaum, and Evans, 1999). Figures 2, 3 and 4 report the median and the 16th and 84th percentiles of the posterior distribution of the impulse responses to a monetary policy shock estimated in the large-scale model, using the full sample. The distribution of the impulse responses encompasses both uncertainty on the parameters and hyperparameters.

INSERT FIGURES FROM 2 TO 4 HERE

A one-standard-deviation (approximately 60 basis points) exogenous increase in the federal funds rate generates a substantial contraction in GDP, employment and all other variables related to economic activity. Monetary aggregates also decrease on impact, indicating strong liquidity effects. Moreover, stock prices decline, the exchange rate appreciates and the yield curve flattens. Prices decrease with a delay. Notice that, with the exception of the CPI, the response of prices does not exhibit the so called price puzzle, i.e. a counterintuitive positive response to a monetary contraction, which is instead typical of VARs with small information sets (Sims, 1992b; Bernanke, Boivin, and Eliasz, 2005; Bańbura, Giannone, and Reichlin, 2010). These responses are all in line with intuition, and hence lend support to our hierarchical procedure. On the other hand, there is no formal way to assess the accuracy of this estimation, since there is

no counterpart of these responses directly observable in the data. This is why we now turn to our second exercise.

In our controlled Monte Carlo experiment, we adopt a medium-scale DSGE model to simulate 500 artificial time series of length of 200 quarters, for the following seven macro variables: output (Y), consumption (C), investment (I), hours worked (H), wages (W), prices (P) and the short-term interest rate (R). For each dataset, we estimate the impulse responses to a monetary policy shock with our hierarchical BVAR model and a flat-prior VAR, and compare these estimates to the true impulse responses of the theoretical model.

The DSGE that we use to simulate the data is identical to Justiniano, Primiceri, and Tambalotti (2010), with the exception that the behavior of the private sector is predetermined with respect to the monetary policy shock, as in Christiano, Eichenbaum, and Evans (2005). This justifies the use of a recursive scheme for the identification of monetary policy shocks in the BVAR and the VAR. Finally, the DSGE is parameterized using the posterior mode of the unknown coefficients, estimated using U.S. data on output growth, consumption growth, investment growth, hours, wage inflation, price inflation and the federal funds rate, as in Justiniano, Primiceri, and Tambalotti (2010). This is a good laboratory to study the question at hand, since it is well known that this class of medium-scale DSGE models fits the data quite well (Smets and Wouters, 2007).

Figure 5 reports the theoretical DSGE impulse responses to a monetary policy shock (solid line), and the average across replications of the median responses using our hierarchical procedure (dashed line) and the flat-prior VAR (dotted line). Both the BVAR and the VAR responses replicate the shape of the true impulse responses quite well. In general, the bias introduced by using an informative prior is not substantially larger than the small sample bias of the flat-prior VAR.[9]

INSERT FIGURE 5 HERE

However, the difference between the average median across replications and the theoretical impulse response, the bias, represents only one dimension of accuracy. In order to take into account also the standard deviation of the errors across replications, we need to look at the average squared error across replications.

More in details, for each replication, we compute the overall error as the difference between the theoretical response and the estimated median response across variables and horizons. Then, for each variable and horizon, we take the average of the squared errors across replications (MSE). Figure 6 reports the ratio between the MSE for the flat-prior VAR and the hierarchical BVAR.

INSERT FIGURE 6 HERE

Such a ratio is greater than one for most variables and horizons, indicating that the hierarchical BVAR yields very substantial accuracy gains. For instance, depending on

[9] We have also computed the impulse responses to a monetary policy shock in the theoretical VAR(5) representation of the DSGE model. These responses are extremely similar to the DSGE responses.

the horizon, the impulse responses of output, consumption, investment, hours and wages based on the BVAR can be about twice as accurate. An important exception is the response of the federal funds rate, which is estimated to be too persistent and to decay too slowly when using informative priors (see figures 5 and 6). Further experimentation reveals that this excessively persistent behavior is due to the sum-of-coefficients prior. While this prior is very important to enhance the forecasting performance of the model, the outcomes in figures 5 and 6 suggest that more sophisticated priors might be needed to discipline the behavior of the model at low frequencies. It is also reasonable to expect that these more sophisticated priors should be based on insights coming from economic theory, since it is well known that the data are less informative about low frequency trends.

6 Conclusion

In this paper, we have studied the problem of how to choose the informativeness of a prior distribution for VAR models. Our approach consists of treating the coefficients of the prior as additional parameters, in the spirit of hierarchical modeling. We have shown that this approach is theoretically grounded, easy to implement, and performs very well both in terms of out-of-sample forecasting, and accuracy in the estimation of impulse response functions. Moreover, it greatly reduces the number and importance of subjective choices in the setting of the prior. In sum, this hierarchical modeling procedure is beneficial for both reduced-form and structural analysis with VARs. Moreover, this approach may prove particularly useful also for the increasingly large literature on DSGE models. It is in fact typical in this literature to validate a theoretical model by comparing its fit and impulse responses to those of VARs.

Appendix A: The Marginal Likelihood for a BVAR with a Conjugate Prior

This appendix derives the functional form of the ML for a BVAR with a conjugate prior. Consider the VAR model of section 1

$$\begin{aligned} y_t &= C + B_1 y_{t-1} + ... + B_p y_{t-p} + \varepsilon_t, \quad t = 1, ..., T \\ \varepsilon_t &\sim N(0, \Sigma), \end{aligned}$$

and rewrite it as

$$\begin{aligned} Y &= X\beta + \epsilon \\ \epsilon &\sim N(0, \Sigma \otimes I_{T-p}), \end{aligned}$$

where $y \equiv [y_{p+1}, ..., y_T]'$, $Y \equiv vec(y)$, $x_t \equiv [1, y'_{t-1}, ..., y'_{t-p}]'$, $x \equiv [x_{p+1}, ..., x_T]'$, $X \equiv I_n \otimes x$, $\varepsilon \equiv [\varepsilon_{p+1}, ..., \varepsilon_T]'$, $\epsilon \equiv vec(\varepsilon)$, $B \equiv [C, B_1, ..., B_p]'$ and $\beta \equiv vec(B)$. Finally, define the number of regressors for each equation by $k \equiv np + 1$.

As in section 2, the prior on (β, Σ) is given by the following Normal-Inverse-Wishart distribution[10]

$$\begin{aligned} \Sigma &\sim IW(\Psi, d) \\ \beta|\Sigma &\sim N(b, \Sigma \otimes \Omega), \end{aligned}$$

where, for simplicity, we are not explicitly conditioning on the hyperparameters b, Ω, Ψ and d.

The un-normalized posterior of (β, Σ) can be obtained by multiplying the prior density by the likelihood function

$$\begin{aligned} p(\beta, \Sigma|Y) = &\left(\frac{1}{2\pi}\right)^{\frac{n(T-p+k)}{2}} |\Sigma|^{-\frac{T-p+k+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \cdot \\ &e^{-\frac{1}{2}\left[\begin{array}{c} (Y - X\beta)'(\Sigma \otimes I_T)^{-1}(Y - X\beta) + \\ + (\beta - b)'(\Sigma \otimes \Omega)^{-1}(\beta - b) \end{array}\right]}. \end{aligned} \tag{6.8}$$

Tedious algebraic manipulations of (6.8) yield the expression

$$\begin{aligned} p(\beta, \Sigma|Y) = &\left(\frac{1}{2\pi}\right)^{\frac{n(T-p+k)}{2}} |\Sigma|^{-\frac{T-p+k+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \cdot \\ &e^{-\frac{1}{2}\left[\begin{array}{c} \left(\beta - \hat{\beta}\right)'\left[X'(\Sigma \otimes I_T)^{-1}X + (\Sigma \otimes \Omega)^{-1}\right]\left(\beta - \hat{\beta}\right) + \\ + \left(\hat{\beta} - b\right)'(\Sigma \otimes \Omega)^{-1}\left(\hat{\beta} - b\right) + \hat{\epsilon}'(\Sigma \otimes I_T)^{-1}\hat{\epsilon} \end{array}\right]}, \end{aligned} \tag{6.9}$$

[10] We are using the following parameterization of the Inverse Wishart density: $p(\Sigma|\Psi, d) = \frac{|\Psi|^{\frac{d}{2}} \cdot |\Sigma|^{-\frac{n+d+1}{2}} \cdot e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)}$.

where $\hat{B} \equiv \left(x'x + \Omega^{-1}\right)^{-1}\left(x'y + \Omega^{-1}b\right)$, $\hat{\beta} \equiv vec\left(\hat{B}\right)$, $\hat{\varepsilon} \equiv y - x\hat{B}$ and $\hat{\epsilon} \equiv vec\left(\hat{\varepsilon}\right)$. It can be shown that (6.9) is the kernel of the following Normal-Inverse-Wishart posterior distribution:

$$\begin{aligned} \Sigma|Y &\sim IW\left(\Psi + \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B} - \hat{b}\right)'\Omega^{-1}\left(\hat{B} - \hat{b}\right), T - p + d\right) \quad (6.10) \\ \beta|\Sigma, Y &\sim N\left(\hat{\beta}, \Sigma \otimes \left(x'x + \Omega^{-1}\right)^{-1}\right), \quad (6.11) \end{aligned}$$

where $\hat{b}$ is a $k \times n$ matrix obtained by reshaping the vector b in such a way that each column corresponds to the prior mean of the coefficients of each equation.

The ML is the integral of the un-normalized posterior:

$$p\left(Y\right) = \int\int p\left(Y|\beta, \Sigma\right) \cdot p\left(\beta|\Sigma\right) \cdot p\left(\Sigma\right) d\beta d\Sigma. \quad (6.12)$$

Let's start with the integral with respect to β. Substituting (6.9) into (6.12) we obtain

$$p\left(Y, \Sigma\right) = \int \left[\begin{array}{c} \left(\frac{1}{2\pi}\right)^{\frac{n(T-p+k)}{2}} |\Sigma|^{-\frac{T-p+k+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \cdot \\ e^{-\frac{1}{2}\left[\begin{array}{c} \left(\beta - \hat{\beta}\right)'\left[X'\left(\Sigma \otimes I_{T-p}\right)^{-1}X + \left(\Sigma \otimes \Omega\right)^{-1}\right]\left(\beta - \hat{\beta}\right) + \\ + \left(\hat{\beta} - b\right)'\left(\Sigma \otimes \Omega\right)^{-1}\left(\hat{\beta} - b\right) + \hat{\epsilon}'\left(\Sigma \otimes I_{T-p}\right)^{-1}\hat{\epsilon} \end{array} \right]} \end{array} \right] d\beta,$$

which can be solved by "completing the squares," yielding

$$\begin{aligned} p\left(Y, \Sigma\right) &= \left(\frac{1}{2\pi}\right)^{\frac{n(T-p)}{2}} |\Sigma|^{-\frac{T-p+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \cdot \\ & e^{-\frac{1}{2}\left[\left(\hat{\beta}-b\right)'\left(\Sigma\otimes\Omega\right)^{-1}\left(\hat{\beta}-b\right)+\hat{\epsilon}'\left(\Sigma\otimes I_{T-p}\right)^{-1}\hat{\epsilon}\right]} \cdot \left|x'x + \Omega^{-1}\right|^{-\frac{n}{2}}. \end{aligned}$$

We are now ready to take the integral with respect to Σ:

$$\begin{aligned} p\left(Y\right) &= \left(\frac{1}{2\pi}\right)^{\frac{n(T-p)}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{1}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \left|x'x + \Omega^{-1}\right|^{-\frac{n}{2}} \\ & \int \left[\begin{array}{c} |\Sigma|^{-\frac{T-p+n+d+1}{2}} e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)} \cdot \\ e^{-\frac{1}{2}\underbrace{\left[\left(\hat{\beta} - b\right)'\left(\Sigma \otimes \Omega\right)^{-1}\left(\hat{\beta} - b\right) + \hat{\epsilon}'\left(\Sigma \otimes I_{T-p}\right)^{-1}\hat{\epsilon}\right]}_{P}} \end{array} \right] d\Sigma. \quad (6.13) \end{aligned}$$

The expression for P can be simplified by using the following property of the *vec* operator:

$$vec\left(A\right)'\left(D \otimes B\right)vec\left(C\right) = tr\left(A'BCD'\right).$$

This yields

$$P = tr\left[\hat{\varepsilon}'\hat{\varepsilon}\Sigma^{-1} + \left(\hat{B} - \hat{b}\right)'\Omega^{-1}\left(\hat{B} - \hat{b}\right)\Sigma^{-1}\right]. \quad (6.14)$$

We can now solve the integral by substituting (6.14) into (6.13), and multiplying and dividing the expression inside the integral by the constant term necessary to obtain the density of an Inverse-Wishart. This results in the following closed-form solution for the ML:

$$\begin{aligned} p(Y) &= \left(\frac{1}{\pi}\right)^{\frac{n(T-p)}{2}} \frac{\Gamma_n\left(\frac{T-p+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)} \cdot \\ & \left|\Omega\right|^{-\frac{n}{2}} \cdot \left|\Psi\right|^{\frac{d}{2}} \cdot \left|x'x + \Omega^{-1}\right|^{-\frac{n}{2}} \cdot \\ & \left|\Psi + \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B} - \hat{b}\right)' \Omega^{-1} \left(\hat{B} - \hat{b}\right)\right|^{-\frac{T-p+d}{2}} \end{aligned}$$

Appendix B: The MCMC Algorithm

This appendix presents the details of the MCMC algorithm that we use to simulate the posterior of the coefficients of the BVAR, including the hyperparameters. We use the following standard Metropolis agorithm:

1. Initialize the hyperparameters γ at their posterior mode, which requires a numerically maximization.

2. Draw a candidate value of the hyperparameters γ^* from a Gaussian proposal distribution, with mean equal to $\gamma^{(j-1)}$ and variance equal to $c \cdot W$, where $\gamma^{(j-1)}$ is the previous draw of γ, W is the inverse Hessian of the negative of the log-posterior of the hyperparameters at the peak, and c is a scaling constant chosen to obtain an acceptance rate of approximately 20 percent.

3. Set
$$\gamma^{(j)} = \begin{cases} \gamma^* \text{ with pr. } \alpha^{(j)} \\ \gamma^{(j-1)} \text{ with pr. } 1 - \alpha^{(j)}, \end{cases}$$
where
$$\alpha^{(j)} = \min\left\{1, \frac{p(\gamma^*|y)}{p(\gamma^{(j-1)}|y)}\right\}$$

4. Draw $\left[\beta^{(j)}, \Sigma^{(j)}\right]$ from $p\left(\beta, \Sigma|y, \gamma^{(j)}\right)$, which is the density of the Normal-Inverse-Wishart distribution in (6.10)-(6.11).

5. Increment j to *j+1* and go to 2.

Appendix C: Factor augmented regression

We consider the following forecasting equation:

$$z_{i,t+h}^h = c_i + \sum_{s=0}^{p_z-1} \alpha_{i,s} z_{i,t-s} + \sum_{k=1}^{r} \lambda_{ik} f_{k,t} + e_{i,t+h}^h$$

where $z^h_{i,t+h}$ denotes the h-steps ahead variable to be forecasted. The predictors $f_{k,t}, k = 1, ..., r$ are common factors extracted from the set of all variables. The lags of the target variable $z_{i,t-s}$ are explicitly used as predictors in order to capture variable specific dynamics. The regression coefficients are allowed to differ across forecast horizons, but the dependence is dropped for notational convenience.

The estimation of the forecasting equation is performed in two steps, as in Stock and Watson (2002a,b). In the first step, the common factors $f_{k,t}$ are estimated by principal components extracted from a large set of 149 predictors. Before extracting the common factors, the data are transformed in order to achieve stationarity and standardized. For details on data definitions and transformations see table 1 and Stock and Watson (2008).

In the second step, the coefficients are estimated by ordinary least squares. Using all the principal components (i.e. by setting r equal to the number of variables 149) would be equivalent to running an OLS regression on all the available regressors. Therefore, as in Stock and Watson (2008), we set $r = 3$ ans $p_z = 4$.

References

Amisano, G., and R. Giacomini (2007): "Comparing density forecasts via weighted likelihood ratio tests," *Journal of Business and Economic Statistics*, 25, 177–190.

Atkeson, A., and L. E. Ohanian (2001): "Are Phillips curves useful for forecasting inflation?," *Quarterly Review*, (Win), 2–11.

Bańbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian VARs," *Journal of Applied Econometrics*, 25(1), 71–92.

Belmonte, M., G. Koop, and D. Korobilis (2011): "Hierarchical shrinkage in time-varying parameter models," MPRA Paper 31827, University Library of Munich, Germany.

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer-Verlag.

Berger, J. O., and L. Berliner (1986): "Robust Bayes and Empirical Bayes Analysis with # -Contaminated Priors," *The Annals of Statistics*, 14, 461–486.

Berger, J. O., W. Strawderman, and T. Dejung (2005): "Posterior Property and Admissibility of Hyperpriors in Normal Hierarchical Models," *The Annals of Statistics*, 33(2), 604–646.

Bernanke, B., J. Boivin, and P. S. Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach," *The Quarterly Journal of Economics*, 120(1), 387–422.

Bloor, C., and T. Matheson (2009): "Real-time conditional forecasts with Bayesian VARs: An application to New Zealand," Reserve Bank of New Zealand Discussion Paper Series DP2009/02, Reserve Bank of New Zealand.

Canova, F. (2007): *Methods for Applied Macroeconomic Research*. Princeton University Press.

Carriero, A., T. Clark, and M. Marcellino (2011): "Bayesian VARs: specification choices and forecast accuracy," Discussion paper.

Carriero, A., G. Kapetanios, and M. Marcellino (2009): "Forecasting exchange rates with a large Bayesian VAR," *International Journal of Forecasting*, 25(2), 400–417.

——— (2010): "Forecasting Government Bond Yields," mimeo, University of London.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1999): "Monetary policy shocks: What have we learned and to what end?," in *Handbook of Macroeconomics*, ed. by J. B. Taylor, and M. Woodford, vol. 1 of *Handbook of Macroeconomics*, chap. 2, pp. 65–148. Elsevier.

——— (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113(1), 1–45.

De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?," *Journal of Econometrics*, 146(2), 318–328.

Del Negro, M., and F. Schorfheide (2004): "Priors from General Equilibrium Models for VARS," *International Economic Review*, 45(2), 643–673.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters (2007): "On the Fit of New Keynesian Models," *Journal of Business & Economic Statistics*, 25, 123–143.

Doan, T., R. Litterman, and C. A. Sims (1984): "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, 3, 1–100.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The Generalized Dynamic Factor Model: identification and estimation," *Review of Economics and Statistics*, 82, 540–554.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004): *Bayesian Data Analysis: Second Edition*. Boca Raton: Chapman and Hall CRC.

Geweke, J. (2001): "Bayesian econometrics and forecasting," *Journal of Econometrics*, 100(1), 11–15.

Geweke, J., and C. Whiteman (2006): *Bayesian Forecasting*chap. 1, pp. 3–80, Handbook of Economic Forecasting. Elsevier.

Giannone, D., M. Lenza, D. Momferatou, and L. Onorante (2010): "Short-Term Inflation Projections: a Bayesian Vector Autoregressive approach," CEPR Discussion Papers 7746, C.E.P.R. Discussion Papers.

Giannone, D., M. Lenza, and L. Reichlin (2008): "Explaining The Great Moderation: It Is Not The Shocks," *Journal of the European Economic Association*, 6(2-3), 621–633.

Jarociski, M., and A. Marcet (2010): "Autoregressions in small samples, priors about observables and initial conditions," Working Paper Series 1263, European Central Bank.

Justiniano, A., G. E. Primiceri, and A. Tambalotti (2010): "Investment shocks and business cycles," *Journal of Monetary Economics*, 57(2), 132–145.

Kadiyala, K. R., and S. Karlsson (1997): "Numerical Methods for Estimation and Inference in Bayesian VAR-Models," *Journal of Applied Econometrics*, 12(2), 99–132.

Knox, T., J. H. Stock, and M. W. Watson (2000): "Empirical Bayes Forecasts of One Time Series Using Many Predictors," Econometric Society World Congress 2000 Contributed Papers 1421, Econometric Society.

Koop, G. (2003): *Bayesian Econometrics*. Wiley.

——— (2011): "Forecasting with Medium and Large Bayesian VARs," *Journal of Applied Econometrics*, forthcoming.

Koop, G., and D. Korobilis (2010): "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics," *Foundations and Trends in Econometrics*, 3(4), 267–358.

Litterman, R. (1979): "Techniques of forecasting using vector autoregressions," Federal Reserve of Minneapolis Working Paper 115.

——— (1980): "A Bayesian Procedure for Forecasting with Vector Autoregression.," Working paper, Massachusetts Institute of Technology, Department of Economics.

——— (1986): "Forecasting With Bayesian Vector Autoregressions – Five Years of Experience," *Journal of Business and Economic Statistics*, 4, 25–38.

Lopes, H. F., A. R. B. Moreira, and A. M. Schmidt (1999): "Hyperparameter estimation in forecast models," *Comput. Stat. Data Anal.*, 29(4), 387–410.

Phillips, P. C. (1995): "Automated Forecasts of Asia-Pacific Economic Activity," Cowles foundation discussion papers, Cowles Foundation for Research in Economics, Yale University.

Phillips, P. C., and W. Ploberger (1994): "Posterior Odds Testing for a Unit Root with Data-Based Model Selection," *Econometric Theory*, 10(3-4), 774–808.

Primiceri, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *Review of Economic Studies*, 72, 821–852.

Robbins, H. (1956): "An Empirical Bayes Approach to Statistics," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 157–163.

Robertson, J. C., and E. W. Tallman (1999): "Vector autoregressions: forecasting and reality," *Economic Review*, (Q1), 4–18.

——— (2001): "Improving Federal-Funds Rate Forecasts in VAR Models Used for Policy Analysis," *Journal of Business & Economic Statistics*, 19(3), 324–30.

Sims, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48(1), 1–48.

Sims, C. A. (1992a): "Bayesian inference for multivariate time series with trend," mimeo, princeton university.

——— (1992b): "Interpreting the macroeconomic time series facts: the effects of monetary policy," *European Economic Review*, 36, 975–1000.

——— (1993): "A Nine-Variable Probabilistic Macroeconomic Forecasting Model," in *Business Cycles, Indicators and Forecasting*, NBER Chapters, pp. 179–212. National Bureau of Economic Research, Inc.

Sims, C. A., and T. Zha (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, 39(4), 949–68.

Smets, F., and R. Wouters (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97(3), 586–606.

Stein, C. (1956): "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," *Proc. Third Berkeley Symp. on Math. Statist. and Prob.*, 1, 197–206.

Stock, J. H., and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 147–162.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economics Statistics*, 20, 147–162.

——— (2008): "Forecasting in Dynamic Factor Models Subject to Structural Instability," in *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, ed. by J. Castle, and N. Shephard. Oxford University Press.

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

Wright, J. H. (2009): "Forecasting US inflation by Bayesian model averaging," *Journal of Forecasting*, 28(2), 131–144.

Tables

Table 1: The description of the database

Variables	Mnemonic	Transf. BVAR	Transf. Factor Model	Small BVAR	Medium BVAR	Large BVAR
Real GDP	RGDP	log	log difference	x	x	x
GDP deflator	PGDP	logs	log difference	x	x	x
Federal Funds Rate	FedFunds	raw	difference	x	x	x
CPI	CPI-ALL	logs	log difference			x
Commodity Price	Com:spotprice(real)	logs	log difference			x
Industrial Production	IP:total	logs	log difference			x
Employment	Emp:total	logs	log difference			x
Unemployment	Emp:services	raw	difference			x
Real Consumption	Cons	logs	log difference		x	x
Real Investment	Inv	logs	log difference		x	
Residential Investment	Res.Inv	logs	log difference			x
Non Residential Investment	NonResInv	logs	log difference			x
Personal Consumption Expenditures, Price Index	PCED	logs	log difference			x
Gross Private Domestic Investment, Price Index	PGPDI	logs	log difference			x
Capacity Utilization	CapacityUtil	raw	difference			x
Consumer expectations	Consumerexpect	raw	difference			x
Hours Worked	Emp.Hours	logs	log difference		x	x
Real compensation per hours	RealComp/Hour	logs	log difference		x	x
One year bond rate	1yrT-bond	raw	difference			x
Five years bond rate	5yrT-bond	raw	difference			x
SP500	S&P500	logs	log difference			x
Effective exchange rate	Exrate:avg	logs	log difference			x
Total reserves	Reservestot	logs	log difference			x
M2	M2	logs	log difference			x

Table 2: BVAR MSFE

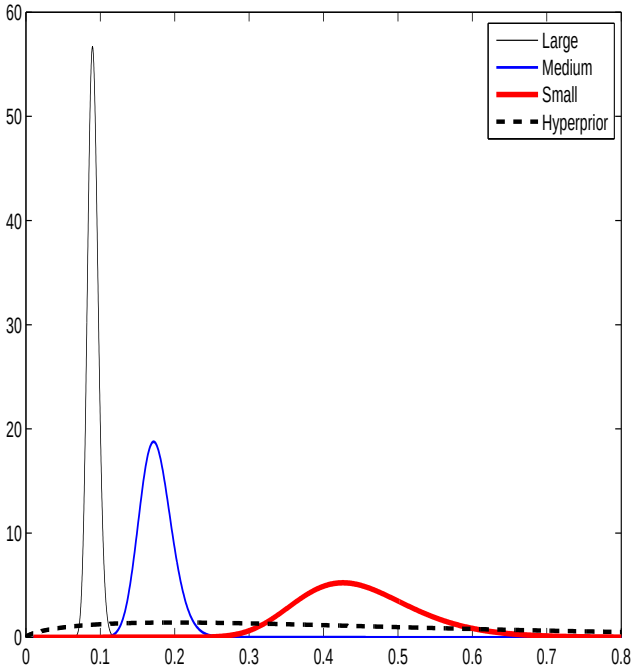
		Small (S)			Medium (M)			Large (L)			Factor M.		RW
Horizons	Variables	VAR	BVAR		VAR	BVAR		VAR	BVAR				
One Quarter	Real GDP	13.57	9.61		19.18	7.97			8.18		7.29		10.23
	GDP Deflator	1.54	1.32		2.27	1.35			1.10		1.14		5.19
	Federal Funds Rates	1.61	1.04		1.83	1.03			1.00		1.25		1.06
One Year	Real GDP	5.39	3.85		11.90	3.42			3.97		3.52		3.98
	GDP Deflator	1.61	1.45		2.22	1.58			0.96		1.01		4.65
	Federal Funds Rates	0.58	0.32		0.56	0.31			0.36		0.32		0.31

Table 3: Average difference of log-scores

		Small (S)			Medium (M)			Large (L)	
Horizons	Variables	vs VAR	vs RW		vs VAR	vs RW		vs VAR	vs RW
One Quarter	Real GDP	0.10	0.06		0.31	0.16			0.17
		(0.04)	(0.05)		(0.05)	(0.06)			(0.06)
	GDP Deflator	0.05	0.74		0.15	0.73			0.81
		(0.03)	(0.09)		(0.05)	(0.09)			(0.09)
	Federal Funds Rates	0.07	0.06		0.10	0.07			0.09
		(0.07)	(0.08)		(0.13)	(0.08)			(0.10)
One Year	Real GDP	0.11	0.00		0.43	0.06			0.03
		(0.07)	(0.09)		(0.12)	(0.09)			(0.13)
	GDP Deflator	0.05	1.00		0.02	0.88			1.18
		(0.10)	(0.33)		(0.22)	(0.36)			(0.30)
	Federal Funds Rates	0.26	0.07		0.27	0.05			-0.03
		(0.07)	(0.07)		(0.12)	(0.09)			(0.12)

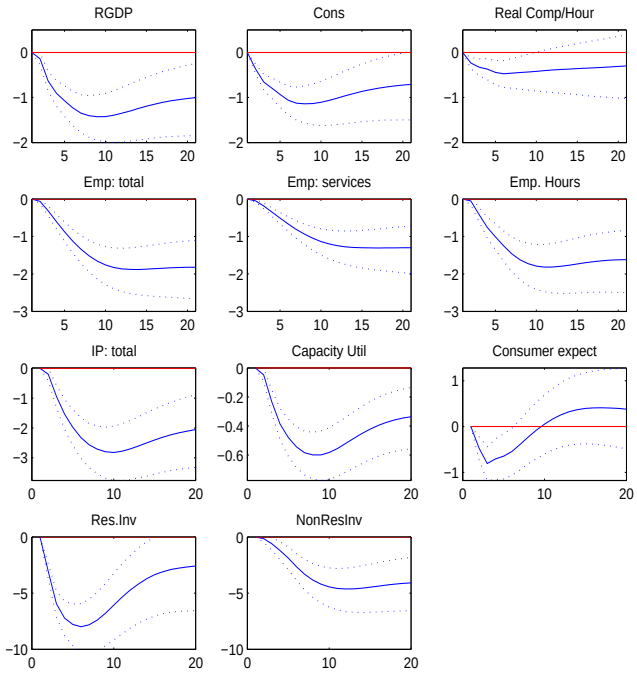
Figures

Figure 1: Posterior distribution of the hyperparameter governing the variance of the Minnesota Prior



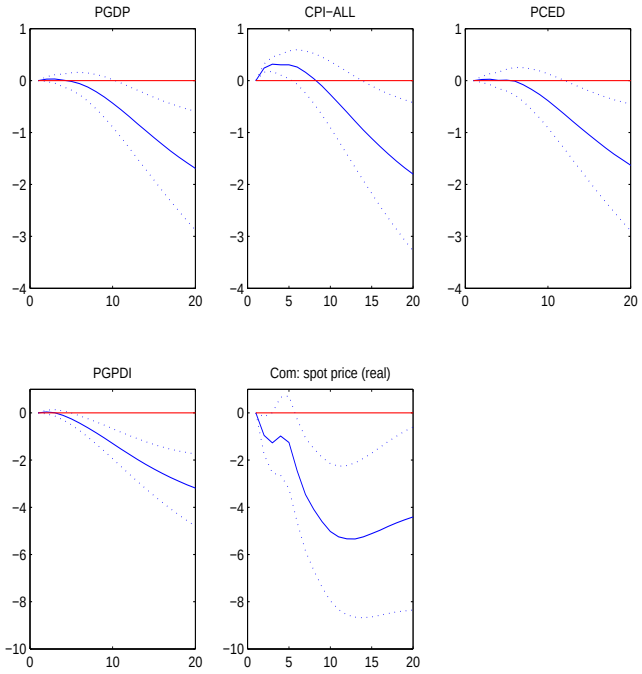
Note: The figure reports the posterior distribution of the hyperparameter λ, the parameter governing the variance of the Minnesota prior in the small, medium, large BVARs, and its prior distribution.

Figure 2: Impulse responses of real variables



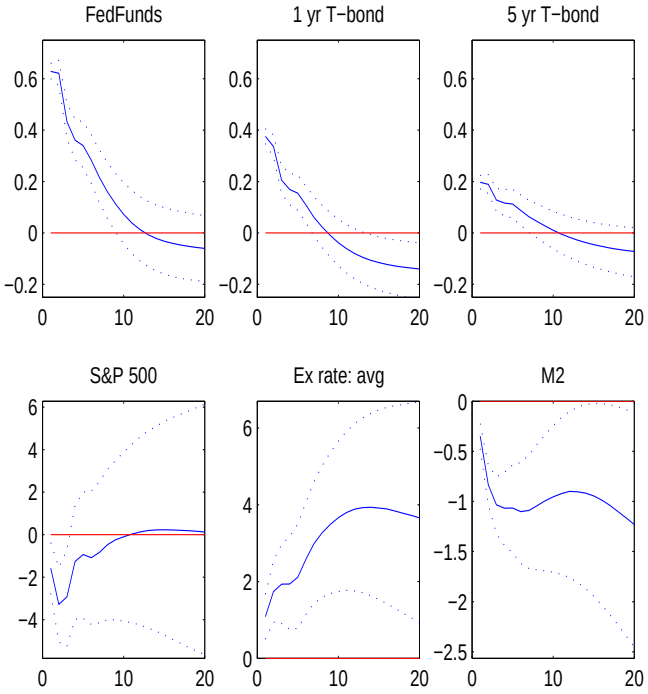
Note: The figure reports the median (solid line) and the 16th and 84th percentiles (dashed lines) of the distribution of the impulse response functions of the large BVAR to a one standard deviation monetary policy shock.

Figure 3: Impulse responses of nominal variables



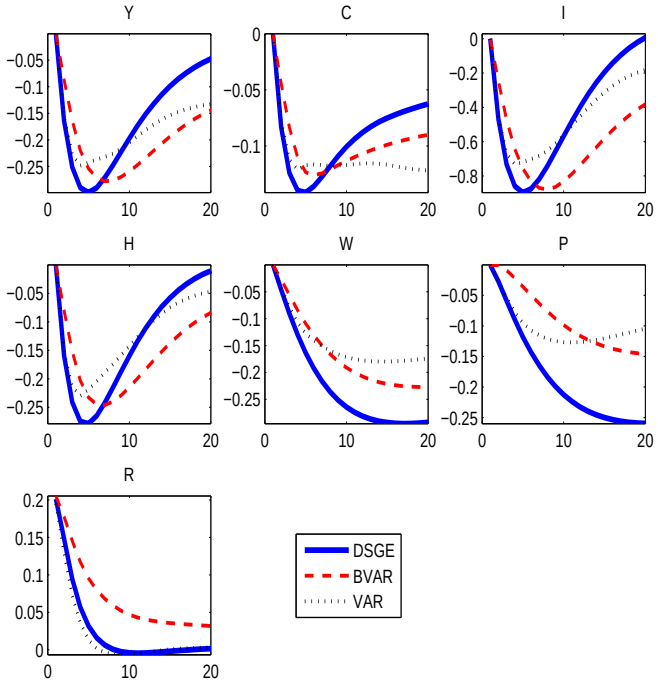
Note: The figure reports the median (solid line) and the 16th and 84th percentiles (dashed lines) of the distribution of the impulse response functions of the large BVAR to a one standard deviation monetary policy shock.

Figure 4: Impulse responses of financial variables



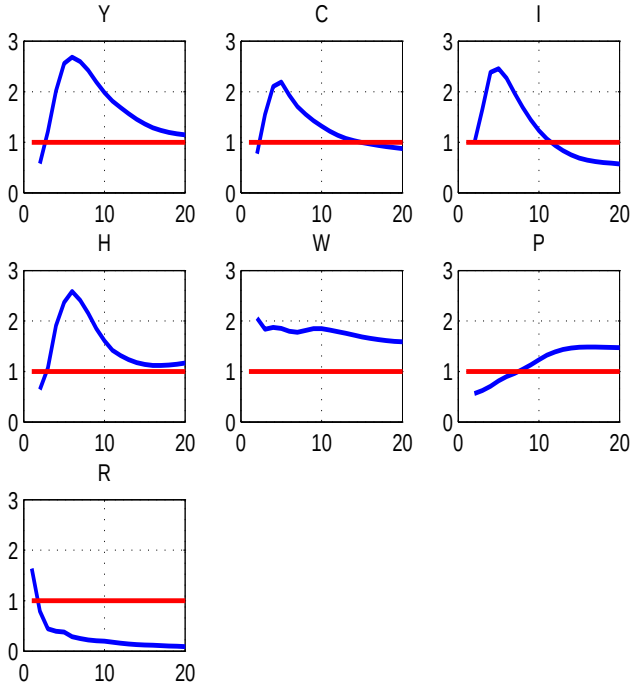
Note: The figure reports the median (solid line) and the 16th and 84th percentiles (dashed lines) of the distribution of the impulse response functions of the large BVAR to a one standard deviation monetary policy shock.

Figure 5: Impulse responses on simulated data



Note: The figure reports the impulse responses to a monetary policy shock in the DSGE model used to generate the data and the median across MonteCarlo replications of the BVAR and the VAR impulse responses.

Figure 6: Ratio of MSE: VAR versus BVAR



Note: The figure reports the ratio of the MSE of the VAR over the MSE of the BVAR. Values larger than one indicate that the MSE of the VAR is larger than that of the BVAR.