# 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: October 2013

**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 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—as well as factor models—and accuracy in the estimation of impulse response functions.

**JEL Codes:** C11, C32, C53, E37
**Keywords:** Forecasting, Bayesian methods, Marginal Likelihood, Hierarchical modeling, impulse responses

## 1 Introduction

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

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

*We thank Liseo Brunero, Guenter Coenen, Gernot Doppelhofer, Raffaella Giacomini, Dimitris Korobilis, Frank Schorfheide, Chris Sims, Raf Wouters and participants in several conferences and seminars for comments and suggestions. Domenico Giannone is grateful to the Actions de Recherche Concertes (contract ARC-AUWB/2010-15/ULB-11) and Giorgio Primiceri to the Alfred P. Sloan Foundation for financial support. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Eurosystem.

where $y_t$ is an $n \times 1$ vector of endogenous variables, $\varepsilon_t$ is an $n \times 1$ vector of exogenous shocks, and $C$, $B_1$,..., $B_p$ and $\Sigma$ 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, 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). 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.

In order to improve the forecasting performance of VAR models, Litterman (1980) and Doan, Litterman, and Sims (1984) have 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. For a more formal illustration of this point 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 the vector $b$ and the matrix $\Omega$ are known, and $\xi$ is a scalar parameter controlling the tightness of the prior information. The conditional posterior of $\beta$ can be obtained by multiplying this prior by the likelihood function. Taking the initial $p$ observations of the sample as given—a standard assumption that we maintain through the entire paper, without explicitly conditioning on these observations—the posterior takes the form

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

where $y \equiv [y_{p+1}, ..., y_T]'$, $x \equiv [x_{p+1}, ..., x_T]'$, $x_t \equiv [1, y'_{t-1}, ..., y'_{t-p}]'$, and $\flat$ is a 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 (i.e. $b \equiv vec\left(\flat\right)$). Notice that, if we choose a lower $\xi$, the prior becomes more informative, the posterior mean of $\beta$ moves towards the prior mean, and the posterior variance falls.

In this context, 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. At time $T$, the distribution of the one-step-ahead forecast is given by

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

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 $\xi$ are likely to be ideal. On the one hand, if $\xi$ is too low and the prior very dogmatic, density forecasts will be very concentrated around $X_{T+1}b$. 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 $\xi$ 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 $X_{T+1}\hat{\beta}\left(\xi\right)$ 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 on the VAR coefficients. 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 over a pre-sample. 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 $p\left(y|\theta\right)$ and a prior distribution $p_\gamma\left(\theta\right)$, where $\theta$ is the vector of the model's parameters and $\gamma$ collects 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 $p_\gamma\left(\theta\right)$ with $p\left(\theta|\gamma\right)$, and evaluating their posterior (Berger, 1985; Koop, 2003). Such a posterior can be obtained by applying Bayes' law, which yields

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

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

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

[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.

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

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:

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

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).

Finally, the hierarchical structure also implies that the unconditional prior for the parameters $\theta$ has a mixed distribution

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

Mixed distributions have generally fatter tails than each of the component distributions $p\left(\theta|\gamma\right)$, a property that robustifies inference. In fact, when the prior has fatter tails than the likelihood, the posterior is less sensitive to extreme discrepancies between prior and likelihood (Berger, 1985; Berger and Berliner, 1986).

### 1.1 Contribution

In this paper, we adopt the 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. Indeed, we derive an expression for the ML showing that it takes duly into account the trade-off between in-sample fit and model complexity.

Because of this feature, the hierarchical BVAR improves over naïve benchmarks and flat-prior VARs, even for small-scale models, for which the optimal shrinkage is low, but not zero. In addition, the hierarchical BVAR outperforms the most popular ad-hoc procedures to select hyperparameters (see Litterman, 1980; Bańbura, Giannone, and Reichlin, 2010). Finally, we find that the forecasting performance of the model typically improves as we include more variables, and it is comparable to that of factor models. This is remarkable because the latter are among the most successful forecasting methods in the literature.

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.

### 1.2 Related literature

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 (see Doan, Litterman and Sims, 1984, Sims and Zha, 1998, and, more recently, Canova, 2007 and Del Negro and Schorfheide, 2012), but seldom formally implemented in this context. Exceptions to this statement are Lopes, Moreira, and Schmidt (1999), who use a hierarchical approach to estimate a small-scale VAR of the Brazilian economy with a Minnesota prior, and Ni and Sun (2003), who exploit an appealing but restrictive hierarchical structure where the hyperparameter controlling the variance of the prior can be integrated out analytically from the prior and the posterior of the VAR coefficients.

Del Negro and Schorfheide (2004) and Del Negro, Schorfheide, Smets, and Wouters (2007) also use the ML to choose the tightness of a prior for VARs derived from the posterior density of a dynamic stochastic general equilibrium model. In the context of

time-varying VARs, the ML 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. 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.

Closer to our framework, Phillips (1995) chooses the hyperparameters of the Minnesota prior for VARs using the asymptotic posterior odds criterion of Phillips and Ploberger (1994), which is also related to the ML. Del Negro and Schorfheide (2004, 2011), Carriero, Kapetanios, and Marcellino (2010) and Carriero, Clark, and Marcellino (2011) have used the ML to select the variance of a Minnesota prior from a grid of possible values. 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; Koop, 2011) and empirical Bayes regressions with large sets of predictors (Knox, Stock, and Watson, 2000; Korobilis, 2013, and references therein).

The rest of the paper is organized as follows. Section 2 and 3 provide some additional 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 can be obtained 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 $\Psi$, $d$, $b$ and $\Omega$ are typically functions of a lower dimensional vector of hyperparameters $\gamma$. 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 (see the surveys of Koop and Korobilis, 2010; Del Negro and Schorfheide, 2011; Karlsson, 2012).[3] Second, the prior (2.3)-(2.4) is conjugate and has the advantage that the ML of the BVAR can be computed in closed form as a function of $\gamma$.

In appendix A, we prove that

$$p(y|\gamma) \propto \underbrace{\left|\left(V_{\varepsilon}^{\text{posterior}}\right)^{-1} V_{\varepsilon}^{\text{prior}}\right|^{\frac{T-p+d}{2}}}_{\text{Fit}} \cdot \underbrace{\prod_{t=p+1}^{T} \left|V_{t|t-1}\right|^{-\frac{1}{2}}}_{\text{Penalty for model complexity}}, \tag{2.5}$$

where $V_{\varepsilon}^{\text{posterior}}$ and $V_{\varepsilon}^{\text{prior}}$ are the posterior and prior means (or modes) of the residual variance, and $V_{t|t-1} \equiv E_{\Sigma}\left[var\left(y_t|y^{t-1}, \Sigma\right)\right]$ is the variance (conditional on $\Sigma$) of the one-step-ahead forecast of $y$, averaged across all possible a-priori realizations of $\Sigma$. While exact closed-form expressions for these objects are provided in the appendix, here we stress that the ML consists of two crucial terms. The first term depends on the *in-sample* fit of the model, and it increases when the posterior residual variance falls relative to the prior variance. Thus, everything else equal, the ML criterion favors hyperparameter values that generate smaller residuals. The second term in (2.5) is instead a penalty for model complexity. This term penalizes models with imprecise *out-of-sample* forecasts due to either large a-priori residual variances or high uncertainty of the parameter estimates. These models have a higher a-priori chance of capturing any possible behavior of the data, while, at the same time, assigning very low probability to all possible outcomes. This feature is the essence of overfitting and is penalized by the ML criterion. Therefore, the ML captures the standard trade-off between model fit and complexity.

The fact that the ML is available in closed form simplifies inference substantially, because it makes it 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. It also 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

[3]Some recent studies have also proposed alternative priors for VARs that do not belong to this family. See, for example, Del Negro and Schorfheide (2004), Villani (2009), Jarociski and Marcet (2010) and Koop (2011).

vector of hyperparameters. Conditional on a value of $\gamma$, the VAR coefficients $[\beta, \Sigma]$ can then be drawn from their posterior, which is Normal-Inverse-Wishart. Appendix B presents the details of this procedure.

We now turn to the empirical application of our methodology.

# 3 Priors and Hyperpriors

This section describes the specific priors that we employ in our empirical analysis. For the sake of comparability with previous studies, we choose the most popular prior densities adopted by the existing literature for the estimation of BVARs in levels. However, it is important to stress that our method is not confined to these priors, but, as mentioned in the previous section, applies more generally to all priors belonging to the class defined by (2.3) and (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 $\Sigma$ (it is equal to $\Psi/(d - n - 1)$). In addition, we take $\Psi$ to be a diagonal matrix with an $n \times 1$ vector $\psi$ on the main diagonal. We treat $\psi$ as an hyperparameter, which differs from the existing literature that has been fixing this parameter using sample information. As for the conditional Gaussian prior for $\beta$, we combine the following prior densities:

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:

$$E\left[(B_s)_{ij} | \Sigma\right] = \begin{cases} 1 & \text{if } i = j \text{ and } s = 1 \\ 0 & \text{otherwise} \end{cases}$$

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

   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 $\lambda$, 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, $\tau_t$ 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 $\mu$ 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 $\delta$ 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 $\lambda$, $\mu$, $\delta$ and $\psi$, which we treat as additional parameters. As hyperpriors for $\lambda$, $\mu$ and $\delta$, 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, the choice of the hyperprior for each element of the vector $\psi/(d-n-1)$, i.e. the prior mean of the main diagonal of $\Sigma$, should be loosely related to the scale of the variables in the model. We pick an Inverse-Gamma with scale and shape equal to $(0.02)^2$ because it seems appropriate for our data expressed in annualized log-terms (see table 1). This hyperprior peaks at approximately $(0.02)^2$ and it is proper, but quite disperse since it does not have either a variance or 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). Another appealing feature of non-flat hyperpriors is that they help stabilize inference when the ML happens to have little curvature with respect to some hyperparameters. For example, we have noticed that this can sometimes occur for the hyperparameters of the sum-of-coefficients or the dummy-initial-observation priors in larger-scale models. This being said, we stress that 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 misspecification. Out-of-sample forecasting reflects both parameter uncertainty and model misspecification 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:[4]

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

[4]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.

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 4), 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 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. $z^h_{i,t+h} = \frac{1}{h}[y_{i,t+h} - y_{i,t}]$. 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)[5] 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.[6] Factor models offer a parsimonious representation for macroeconomic variables, while retaining the salient features of the data that

[5]Precisely, the flat-prior VAR is estimated using a standard uninformative reference prior proportional to $|\Sigma|^{-(n+1)/2}$, which makes the posterior distribution of $\beta$ equivalent to the sampling distribution of its OLS estimator.

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

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. Finally, in a later subsection, we compare the forecasting performance of our hierarchical BVAR to more heuristic procedures for the choice of hyperparameters.

### 4.1 Point forecasts

Table 2 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 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 the one-year horizon. The improvement of BVARs over the RW, which is 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 naïve 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.

### 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 the 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. We measure the log-predictive score using a Gaussian approximation of the predictive density for all models.

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 more accurate than 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.[7]

INSERT TABLE 3 HERE

Table 3 makes clear that the BVAR forecasts outperform 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 parsimonious 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 $\Sigma$ 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 $\lambda$ governing the overall standard deviation of the Minnesota 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 $\lambda$. These posteriors are plotted in figure 1, along with the hyperprior.

[7] Notice that the associated t-statistic corresponds to the statistic 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.

Notice that, in line with intuition, the posterior mode (and variance) of $\lambda$ 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

### 4.4 Comparison with alternative methods

Given the good forecasting performance of our inference-based methodology for choosing the hyperparameters (as good as that of factor models), a section discussing the relative performance of alternative methods seems warranted. However, formal alternatives to the marginal likelihood are absent in the literature. For instance, the Bayesian or the Akaike information criteria cannot be adopted because their penalization for model complexity only involves the number of parameters, and does not depend on the value of the hyperparameters. As a consequence, both of these criteria would favor models with loose priors that maximize the model in-sample fit.

An informal method to choose the hyperparameters is to maximize the model forecasting performance over a pre-sample, as in Litterman (1980). An alternative possibility is to control for over-fitting by targeting a desired in-sample fit, as in Bańbura, Giannone, and Reichlin (2010). These heuristic procedures can be interpreted as rough empirical Bayes estimators, and their ad-hoc nature might partly explain why Bayesian VARs have encountered a number of opponents, especially among non-Bayesian researchers. These approaches obviously raise a number of questions: what is the right size of the pre-sample and the forecasting horizon? Should we minimize the MSFE or control for the in-sample fit of all the variables or just those of interest? Moreover, these procedures make it hard to conduct inference incorporating hyperparameter uncertainty. Despite these limitations, these are the most popular approaches in the literature, and we have compared them to our methodology.

Concerning the first method, we have repeated our forecasting experiment by choosing at each point in time the hyperparameters that maximize the past forecasting ability of the VAR. In particular, to follow Litterman (1980) as close as possible, the measure of out-of-sample forecasting performance is the Theil-U statistic, computed over the previous 5 years, and averaged across variables and forecasting horizons (1 to 4). As for the second method, we have replicated Bańbura, Giannone, and Reichlin (2010) by setting the hyperparameters in the medium and large BVARs to match the average in-sample fit of the small VAR with flat priors.[8]

Table 4 reports the MSFE of the Litterman (1980) (LIT) and Bańbura, Giannone, and Reichlin (2010) (BGR) methods relative to ours. A value bigger than one indicates that our method outperforms the alternatives. The general finding is that the performance of these two approaches is similar, and considerably worse than our methodology, particularly for forecasting inflation.

INSERT TABLE 4 HERE

[8]Bańbura, Giannone, and Reichlin (2010) define the in-sample fit as the percentage deviation of the in-sample MSFE from the MSFE of the no-change forecast.

Finally, note that some authors do not even perform an informal search for the optimal hyperparameters, but simply use values from previous studies. For example, a common choice are the hyperparameters of Sims and Zha (1998), which are also the values around which we center our hyperpriors. We have experimented with these fixed hyperparameters and, quite interestingly, have found that they improve over the heuristic procedures of Litterman (1980) and Bańbura, Giannone, and Reichlin (2010), in our empirical application. In fact, as shown in table 4 (columns SZ), the MSFE is only up to 20 percent worse than our method for the small and medium BVAR, and comparable to our method, if not slightly better, for the large BVAR.

This result suggests that the overall tightness implied by the fixed hyperparameters of Sims and Zha (1998) is too low for the small- and medium-scale VARs, while it is approximately "correct" for the large VAR. In order to support this interpretation, we have also experimented with an "extra-large" VAR model with 35 variables, for which we would expect the priors of Sims and Zha (1998) to be too loose.[9] In line with this intuition, the forecasting performance of the BVAR of Sims and Zha (1998) deteriorates relative to ours. In particular, it becomes marginally worse (between 2% and 4% across variables) at the one-quarter horizon, and sensibly worse at the one-year horizon, especially for the federal funds rate and the GDP deflator (with an 11% and 29% higher MSFE, respectively).

In addition, it is worth noticing that the specific values of the hyperparameters used by Sims and Zha (1998) are not guaranteed to work well for other applications—possibly outside the range of US macroeconomic time series—and cannot be applied to different priors. On the contrary, the main appeal of our methodology is that it can be used in a wide range of models and applications, requiring little human judgement in the search for reasonable ranges of hyperparameters. Consequently, there is also less need for extensive robustness checks that characterize empirical works using more ad-hoc methodologies.

# 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 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.

[9]The extra-large model is constructed by adding the following 13 variables to the large VAR: real durable consumption, total housing starts, purchasing managers index (PMI), new orders of consumer goods and materials, real exports, real imports, exports price index, imports price index, unemployment rate, Moody's AAA corporate bond yields, Moody's BAA corporate bond yields, business loans and consumer credit outstanding.

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 (on this point, see Sims, 1992b; Bernanke, Boivin, and Eliasz, 2005; Bańbura, Giannone, and Reichlin, 2010).

For comparison, figures 2, 3 and 4 also report the corresponding quantiles of the distribution of impulse responses of a VAR estimated with flat priors (grey shaded areas). It is evident that these error bands reflect a considerable amount of estimation uncertainty and their width does not allow any meaningful conclusions about the effects of an exogenous monetary tightening. In addition, even when initially significant, these impulse responses tend to revert to zero at a very fast pace, a symptom of a possibly severe small-sample bias toward stationarity. These results 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.[10]

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

[10] 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 (on this point, see, for example, Del Negro and Schorfheide, 2004; Villani, 2009), 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 variety of commonly used prior distributions 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.

# A The Marginal Likelihood of BVARs with Conjugate Priors

This appendix derives an analytical expression for the ML of BVARs with conjugate priors (possibly implemented using dummy observations), and proves the fit-complexity trade-off result that we have stated in (2.5) in the main text.

## A.1 Analytical derivation of the ML

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 $y \equiv [y_{p+1}, ..., y_T]'$, $Y \equiv vec(y)$, $x_t \equiv [1, y'_{t-1}, ..., y'_{t-p}]'$, $X_t \equiv I_n \otimes x'_t$, $x \equiv [x_{p+1}, ..., x_T]'$, $X \equiv I_n \otimes x$, $\varepsilon \equiv [\varepsilon_{p+1}, ..., \varepsilon_T]'$, $\epsilon \equiv vec(\varepsilon)$, $B \equiv [C, B_1, ..., B_p]'$ 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 $(\beta, \Sigma)$ is given by the following Normal-Inverse-Wishart distribution[11]

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

where, for simplicity, we are not explicitly conditioning on the hyperparameters $b$, $\Omega$, $\Psi$ and $d$.

The unnormalized posterior of $(\beta, \Sigma)$ can be obtained by multiplying the prior density by the likelihood function. If we condition on the initial $p$ observations of the sample, which is a standard assumption, we obtain:

$$\begin{aligned} p(\beta, \Sigma|Y) = &\left(\frac{1}{2\pi}\right)^{\frac{n(T-p+k)}{2}} |\Sigma|^{-\frac{T-p+k+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \cdot \\ &e^{-\frac{1}{2}\left[\begin{array}{c} (Y - X\beta)'(\Sigma \otimes I_T)^{-1}(Y - X\beta) + \\ + (\beta - b)'(\Sigma \otimes \Omega)^{-1}(\beta - b) \end{array}\right]}. \end{aligned} \tag{A.6}$$

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

Tedious algebraic manipulations of (A.6) yield the expression

$$p(\beta,\Sigma|Y) = \left(\frac{1}{2\pi}\right)^{\frac{n(T-p+k)}{2}} |\Sigma|^{-\frac{T-p+k+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}}{2^{\frac{nd}{2}}\cdot\Gamma_n\left(\frac{d}{2}\right)} \cdot e^{-\frac{1}{2}\left[\begin{array}{c}\left(\beta-\hat{\beta}\right)'\left[X'\left(\Sigma\otimes I_T\right)^{-1}X+\left(\Sigma\otimes\Omega\right)^{-1}\right]\left(\beta-\hat{\beta}\right)+\\ +\left(\hat{\beta}-b\right)'\left(\Sigma\otimes\Omega\right)^{-1}\left(\hat{\beta}-b\right)+\hat{\epsilon}'\left(\Sigma\otimes I_T\right)^{-1}\hat{\epsilon}\end{array}\right]} \quad \text{(A.7)}$$

where $\hat{B}\equiv\left(x'x+\Omega^{-1}\right)^{-1}\left(x'y+\Omega^{-1}\flat\right)$, $\hat{\beta}\equiv vec\left(\hat{B}\right)$, $\hat{\varepsilon}\equiv y-x\hat{B}$, $\hat{\epsilon}\equiv vec\left(\hat{\varepsilon}\right)$, and $\flat$ 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 (i.e. $b\equiv vec\left(\flat\right)$). It can then be shown that (A.7) is the kernel of the following Normal-Inverse-Wishart posterior distribution:

$$\Sigma|Y \sim IW\left(\Psi+\hat{\varepsilon}'\hat{\varepsilon}+\left(\hat{B}-\flat\right)'\Omega^{-1}\left(\hat{B}-\flat\right),T-p+d\right) \quad \text{(A.8)}$$

$$\beta|\Sigma,Y \sim N\left(\hat{\beta},\Sigma\otimes\left(x'x+\Omega^{-1}\right)^{-1}\right). \quad \text{(A.9)}$$

The ML is the integral of the unnormalized posterior:

$$p(Y)=\int\int p(Y|\beta,\Sigma)\cdot p(\beta|\Sigma)\cdot p(\Sigma)\,d\beta d\Sigma. \quad \text{(A.10)}$$

Let's start with the integral with respect to $\beta$. Substituting (A.7) into (A.10) we obtain

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

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

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

We are now ready to take the integral with respect to $\Sigma$:

$$p(Y) = \left(\frac{1}{2\pi}\right)^{\frac{n(T-p)}{2}}|\Omega|^{-\frac{n}{2}}|\Psi|^{\frac{d}{2}}\frac{1}{2^{\frac{nd}{2}}\cdot\Gamma_n\left(\frac{d}{2}\right)}\left|x'x+\Omega^{-1}\right|^{-\frac{n}{2}} \int\left[\begin{array}{c}|\Sigma|^{-\frac{T-p+n+d+1}{2}}e^{-\frac{1}{2}tr\left(\Psi\Sigma^{-1}\right)}\cdot\\ e^{-\frac{1}{2}\underbrace{\left[\left(\hat{\beta}-b\right)'\left(\Sigma\otimes\Omega\right)^{-1}\left(\hat{\beta}-b\right)+\hat{\epsilon}'\left(\Sigma\otimes I_{T-p}\right)^{-1}\hat{\epsilon}\right]}_{P}}\end{array}\right]d\Sigma \quad \text{(A.11)}$$

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

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

This yields

$$P = tr\left[\hat{\varepsilon}'\hat{\varepsilon}\Sigma^{-1} + \left(\hat{B} - \flat\right)'\Omega^{-1}\left(\hat{B} - \flat\right)\Sigma^{-1}\right]. \tag{A.12}$$

We can now solve the integral by substituting (A.12) into (A.11), and multiplying and dividing the expression inside the integral by the constant term necessary to obtain the density of an Inverse-Wishart. These operations result in the following closed-form solution for the ML:

$$\begin{aligned} p\,(Y) \;=\; & \left(\frac{1}{\pi}\right)^{\frac{n(T-p)}{2}} \frac{\Gamma_n\left(\frac{T-p+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)} \cdot \\ & |\Omega|^{-\frac{n}{2}} \cdot |\Psi|^{\frac{d}{2}} \cdot \left|x'x + \Omega^{-1}\right|^{-\frac{n}{2}} \cdot \\ & \left|\Psi + \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B} - \flat\right)'\Omega^{-1}\left(\hat{B} - \flat\right)\right|^{-\frac{T-p+d}{2}}. \end{aligned} \tag{A.13}$$

## A.2 Numerical issues

For large systems (in our case, for the medium- and the large-scale models), (A.13) is numerically unstable and it is convenient to replace it with the equivalent expression

$$\begin{aligned} p\,(Y) \;=\; & \left(\frac{1}{\pi}\right)^{\frac{n(T-p)}{2}} \frac{\Gamma_n\left(\frac{T-p+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)} \cdot \\ & |\Psi|^{-\frac{T-p}{2}} \cdot \left|D_\Omega' x'x D_\Omega + I_k\right|^{-\frac{n}{2}} \cdot \\ & \left|I_n + D_\Psi'\left[\hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B} - \flat\right)'\Omega^{-1}\left(\hat{B} - \flat\right)\right]D_\Psi\right|^{-\frac{T-p+d}{2}}, \end{aligned} \tag{A.14}$$

where $D_\Omega D_\Omega' = \Omega$ and $D_\Psi D_\Psi' = \Psi^{-1}$. The last two determinants can be computed as the product of one plus the eigenvalues of $D_\Omega' x'x D_\Omega$ and $D_\Psi'\left[\hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B} - \flat\right)'\Omega^{-1}\left(\hat{B} - \flat\right)\right]D_\Psi$ respectively, which is numerically stable.

## A.3 The ML with dummy observations

It is common in the literature to implement some conjugate priors using dummy observations (e.g. the sum-of-coefficients and the dummy-initial-observation priors). In this case, the ML is given by $p\,(Y^{\oplus})\,/p\,(Y^*)$, where $p\,(\cdot)$ is the function (A.13) or (A.14), $Y^*$ denotes the dummy observations, and $Y^{\oplus}$ is the extended set of data, consisting of $Y$ and $Y^*$.

## A.4 Proof of the fit-complexity trade-off result

In order to prove (2.5), rewrite (A.13) as

$$\begin{aligned} p(Y) &= \left(\frac{1}{\pi}\right)^{\frac{n(T-p)}{2}} \frac{\Gamma_n\left(\frac{T-p+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)} \cdot \left(\frac{T-p+d+n+1}{d+n+1}\right)^{-\frac{T-p+d}{2}} \\ &\quad |\Omega|^{-\frac{n}{2}} \cdot |\Psi|^{-\frac{T-p}{2}} \cdot \left|x'x+\Omega^{-1}\right|^{-\frac{n}{2}} \cdot \\ &\quad \left| \left( \frac{\Psi + \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B}-\flat\right)' \Omega^{-1} \left(\hat{B}-\flat\right)}{T-p+d+n+1} \right)^{-1} \frac{\Psi}{d+n+1} \right|^{\frac{T-p+d}{2}} . \end{aligned} \quad (A.15)$$

Define $x^t \equiv [x_{p+1}, ..., x_t]'$ and notice that $x^{t\prime}x^t$ can be written recursively as $x^{t\prime}x^t = x^{t-1\prime}x^{t-1} + x_t x_t'$. The matrix determinant lemma (Harville, 1997) implies that $\left|x^{t\prime}x^t + \Omega^{-1}\right|$ can also be expressed recursively as

$$\left|x^{t\prime}x^t + \Omega^{-1}\right| = \left|x^{t-1\prime}x^{t-1} + \Omega^{-1}\right| \cdot \left(1 + x_t' \left(x^{t-1\prime}x^{t-1} + \Omega^{-1}\right)^{-1} x_t\right). \quad (A.16)$$

The iteration of (A.16), starting from the initial value $\left|\Omega^{-1}\right|$, allows to derive

$$\left|x'x + \Omega^{-1}\right| = \left|\Omega^{-1}\right| \prod_{t=p+1}^{T} \left(1 + x_t' \left(x^{t-1\prime}x^{t-1} + \Omega^{-1}\right)^{-1} x_t\right).$$

If we substitute this last expression into (A.15), we obtain

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

which, using the properties of the determinants and Kronecker products, can be rewritten as

$$\begin{aligned} p(Y) &= \left(\frac{1}{\pi}\right)^{\frac{n(T-p)}{2}} \frac{\Gamma_n\left(\frac{T-p+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)} \cdot \left(\frac{T-p+d+n+1}{d+n+1}\right)^{-\frac{T-p+d}{2}} \\ &\quad \prod_{t=p+1}^{T} \left|\Psi \otimes \left(1 + x_t' \left(x^{t-1\prime}x^{t-1} + \Omega^{-1}\right)^{-1} x_t\right)\right|^{-\frac{1}{2}} \\ &\quad \left| \left( \frac{\Psi + \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{B}-\flat\right)' \Omega^{-1} \left(\hat{B}-\flat\right)}{T-p+d+n+1} \right)^{-1} \frac{\Psi}{d+n+1} \right|^{\frac{T-p+d}{2}} . \end{aligned}$$

Finally, notice that

$$\begin{aligned} E_\Sigma\left[var\left(y_t|y^{t-1},\Sigma\right)\right] &= E_\Sigma\left[X_t\left(\Sigma\otimes\left(x^{t-1\prime}x^{t-1}+\Omega^{-1}\right)^{-1}\right)X_t'+\Sigma\right] \\ &= E_\Sigma\left[\Sigma\otimes\left(1+x_t'\left(x^{t-1\prime}x^{t-1}+\Omega^{-1}\right)^{-1}x_t\right)\right] \\ &= \frac{\Psi}{d-n-1}\otimes\left(1+x_t'\left(x^{t-1\prime}x^{t-1}+\Omega^{-1}\right)^{-1}x_t\right), \end{aligned}$$

where $E_\Sigma$ denotes the expectation operator with respect to $\Sigma$. We can now express the ML as

$$\begin{aligned} p\left(Y\right) &= \text{const}\cdot\left|\left(V_\varepsilon^{\text{posterior}}\right)^{-1}V_\varepsilon^{\text{prior}}\right|^{\frac{T-p+d}{2}}\cdot\prod_{t=p+1}^{T}\left|V_{t|t-1}\right|^{-\frac{1}{2}} \\ \text{const} &\equiv \left(\frac{1}{\pi}\right)^{\frac{n(T-p)}{2}}\frac{\Gamma_n\left(\frac{T-p+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)}\cdot\frac{(d-n-1)^{\frac{d}{2}}}{(T-p+d-n-1)^{\frac{T-p+d}{2}}} \\ V_{t|t-1} &\equiv E_\Sigma\left[var\left(y_t|y^{t-1},\Sigma\right)\right]=\frac{\Psi}{d-n-1}\otimes\left(1+x_t'\left(x^{t-1\prime}x^{t-1}+\Omega^{-1}\right)^{-1}x_t\right) \\ V_\varepsilon^{\text{prior}} &\equiv E\left[\Sigma\right]=\frac{\Psi}{d-n-1} \\ V_\varepsilon^{\text{posterior}} &\equiv E\left[\Sigma|Y\right]=\frac{\Psi+\hat{\varepsilon}'\hat{\varepsilon}+\left(\hat{B}-\flat\right)'\Omega^{-1}\left(\hat{B}-\flat\right)}{T-p+d-n-1}, \end{aligned}$$

where $V_\varepsilon^{\text{prior}}$ and $V_\varepsilon^{\text{posterior}}$ are the prior and posterior means of the residual variance, and their analytical expressions follow from the properties of the Inverse-Wishart distribution.

# 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 Metropolis algorithm:

1. Initialize the hyperparameters $\gamma$ at their posterior mode, which requires a numerically maximization.

2. Draw a candidate value of the hyperparameters $\gamma^*$ 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 $\gamma$, $W$ 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

$$\alpha^{(j)} = \min\left\{1, \frac{p\left(\gamma^{*}|y\right)}{p\left(\gamma^{(j-1)}|y\right)}\right\}$$

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 (A.8)-(A.9).

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

# 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 $z_{i,t+h}^{h}$ denotes the $h$-step 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 $r = 3$ and $p_z = 4$.

# 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, Federal Reserve Bank of Minneapolis*, (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.

——— (2011): "Bayesian Macroeconometrics," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, pp. 293–389. Oxford University Press.

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," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, chap. 1, pp. 3–80. 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.

Harville, D. (1997): *Matrix Algebra from a Statistician's Perspective*. Springer Verlag.

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.

Karlsson, S. (2012): "Forecasting with Bayesian Vector Autoregressions," Working Papers 2012:12, Orebro University, Swedish Business School.

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.

Korobilis, D. (2013): "Hierarchical Shrinkage Priors for Dynamic Regressions with Many Predictors," *International Journal of Forecasting*, 29, 43–59.

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.

Ni, S., and D. Sun (2003): "Noninformative priors and frequentist risks of bayesian estimators of vector-autoregressive models," *Journal of Econometrics*, 115(1), 159–197.

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.

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 | 4·logs | log-diff. | x | x | x |
| GDP deflator | PGDP | 4·logs | log-diff. | x | x | x |
| Federal Funds Rate | FedFunds | raw | diff. | x | x | x |
| Consumer Price Index | CPI-ALL | 4·logs | log-diff. | | | x |
| Commodity Price | Com:spotprice(real) | 4·logs | log-diff. | | | x |
| Industrial Production | IP:total | 4·logs | log-diff. | | | x |
| Employment | Emp:total | 4·logs | log-diff. | | | x |
| Employment in the services sector | Emp:services | 4·logs | log-diff. | | | x |
| Real Consumption | Cons | 4·logs | log-diff. | | x | x |
| Real Investment | Inv | 4·logs | log-diff. | | x | |
| Real Residential Investment | Res.Inv | 4·logs | log-diff. | | | x |
| Non Residential Investment | NonResInv | 4·logs | log-diff. | | | x |
| Personal Consumption Expenditures, Price Index | PCED | 4·logs | log-diff. | | | x |
| Gross Private Domestic Investment, Price Index | PGPDI | 4·logs | log-diff. | | | x |
| Capacity Utilization | CapacityUtil | raw | diff. | | | x |
| Consumer expectations | Consumerexpect | raw | diff. | | | x |
| Hours Worked | Emp.Hours | 4·logs | log-diff. | | x | x |
| Real compensation per hours | RealComp/Hour | 4·logs | log-diff. | | x | x |
| One year bond rate | 1yrT-bond | raw | diff. | | | x |
| Five years bond rate | 5yrT-bond | raw | diff. | | | x |
| SP500 | S&P500 | 4·logs | log-diff. | | | x |
| Effective exchange rate | Exrate:avg | 4·logs | log-diff. | | | x |
| M2 | M2 | 4·logs | log-diff. | | | x |

Table 2: MSFE of point forecasts

| Horizons | Variables | Small (S) VAR | Small (S) BVAR | Medium (M) VAR | Medium (M) BVAR | Large (L) VAR | Large (L) BVAR | Factor M. | RW |
|---|---|---|---|---|---|---|---|---|---|
| One Quarter | Real GDP | 13.49 | 9.61 | 19.15 | 7.97 | | 8.18 | 7.29 | 10.23 |
| | GDP Deflator | 1.53 | 1.32 | 2.26 | 1.35 | | 1.10 | 1.14 | 5.19 |
| | Federal Funds Rates | 1.61 | 1.04 | 1.82 | 1.03 | | 1.00 | 1.25 | 1.06 |
| One Year | Real GDP | 5.40 | 3.85 | 12.10 | 3.42 | | 3.97 | 3.52 | 3.98 |
| | GDP Deflator | 1.61 | 1.45 | 2.25 | 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 |

**Note**: The table reports the mean squared forecast errors of the BVARs and the competing models (VAR: flat-prior VAR, RW: Random Walk in levels with drift, Factor M.: factor augmented regression), for each variable and horizon. The evaluation sample is 1975Q1 - 2008Q4 for the one-quarter-ahead forecasts and 1975Q4 - 2008Q4 for the one-year-ahead forecasts.

Table 3: Average difference of log-scores

| Horizons | Variables | Small (S) vs VAR | Small (S) vs RW | Medium (M) vs VAR | Medium (M) vs RW | Large (L) vs VAR | Large (L) vs RW |
|---|---|---|---|---|---|---|---|
| One Quarter | Real GDP | 0.10 | 0.06 | 0.30 | 0.16 | | 0.17 |
| | | (0.04) | (0.05) | (0.05) | (0.06) | | (0.06) |
| | GDP Deflator | 0.04 | 0.74 | 0.15 | 0.73 | | 0.81 |
| | | (0.03) | (0.09) | (0.06) | (0.09) | | (0.09) |
| | Federal Funds Rates | 0.08 | 0.06 | 0.11 | 0.07 | | 0.09 |
| | | (0.06) | (0.08) | (0.10) | (0.08) | | (0.10) |
| One Year | Real GDP | 0.10 | 0.00 | 0.40 | 0.06 | | 0.03 |
| | | (0.07) | (0.09) | (0.12) | (0.09) | | (0.13) |
| | GDP Deflator | 0.05 | 1.00 | 0.01 | 0.88 | | 1.18 |
| | | (0.10) | (0.33) | (0.22) | (0.36) | | (0.30) |
| | Federal Funds Rates | 0.28 | 0.07 | 0.25 | 0.05 | | -0.03 |
| | | (0.08) | (0.07) | (0.10) | (0.09) | | (0.12) |

**Note**: The table 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. The HAC estimate of the standard deviation of the difference between the log-predictive scores of the BVARs and the competing models is reported in parenthesis. The evaluation sample is 1975Q1 - 2008Q4 for the one-quarter-ahead forecasts and 1975Q4 - 2008Q4 for the one-year-ahead forecasts.

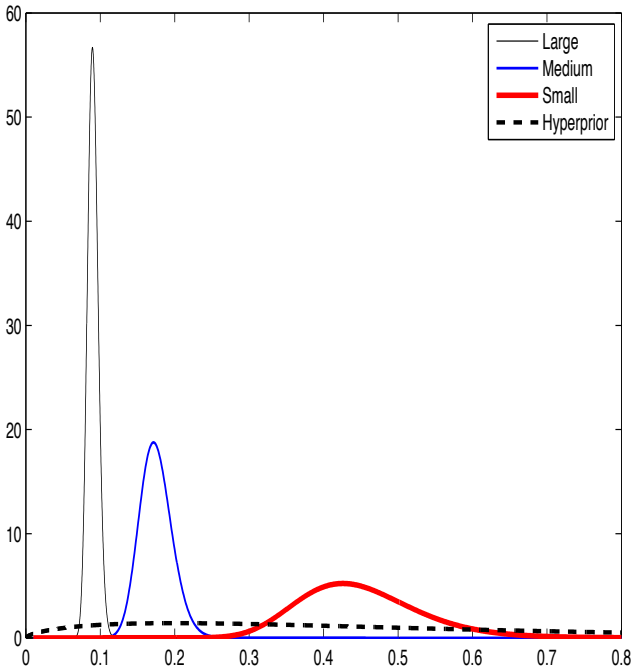
Table 4: MSFE of alternative methods relative to hierarchical model

| Horizons | Variables | Small (S) LIT | Small (S) BGR | Small (S) SZ | Medium (M) LIT | Medium (M) BGR | Medium (M) SZ | Large (L) LIT | Large (L) BGR | Large (L) SZ |
|---|---|---|---|---|---|---|---|---|---|---|
| One Quarter | Real GDP | 1.04 | | 1.02 | 1.19 | 1.09 | 1.06 | 1.12 | 1.09 | 0.96 |
| | GDP Deflator | 1.87 | | 1.09 | 1.67 | 1.44 | 1.11 | 1.46 | 1.97 | 0.97 |
| | Federal Funds Rates | 1.32 | | 1.01 | 1.01 | 0.99 | 0.99 | 1.19 | 1.02 | 0.98 |
| One Year | Real GDP | 1.16 | | 1.10 | 1.15 | 1.17 | 1.14 | 1.32 | 0.97 | 0.87 |
| | GDP Deflator | 1.61 | | 1.23 | 1.57 | 1.62 | 1.21 | 1.55 | 2.73 | 1.04 |
| | Federal Funds Rates | 1.13 | | 0.97 | 1.13 | 1.03 | 1.00 | 1.03 | 0.83 | 0.92 |

**Note**: The table reports the MSFE of three alternative methods to select the tightness of the prior distributions relative to the MSFE of the hierarchical model. Numbers bigger than one indicate that the MSFE of the alternative method (LIT: method described in Litterman (1980); BGR: method described in Bańbura, Giannone, and Reichlin (2010); SZ: fixed hyperparameters in Sims and Zha (1998)) is bigger than the corresponding MSFE of the hierarchical model. The evaluation sample is 1975Q1 - 2008Q4 for the one-quarter-ahead forecasts and 1975Q4 - 2008Q4 for the one-year-ahead forecasts. By construction, the MSFE of the BGR method for the small-scale model (not reported here, see table 2) is identical to that of the flat-prior VAR.

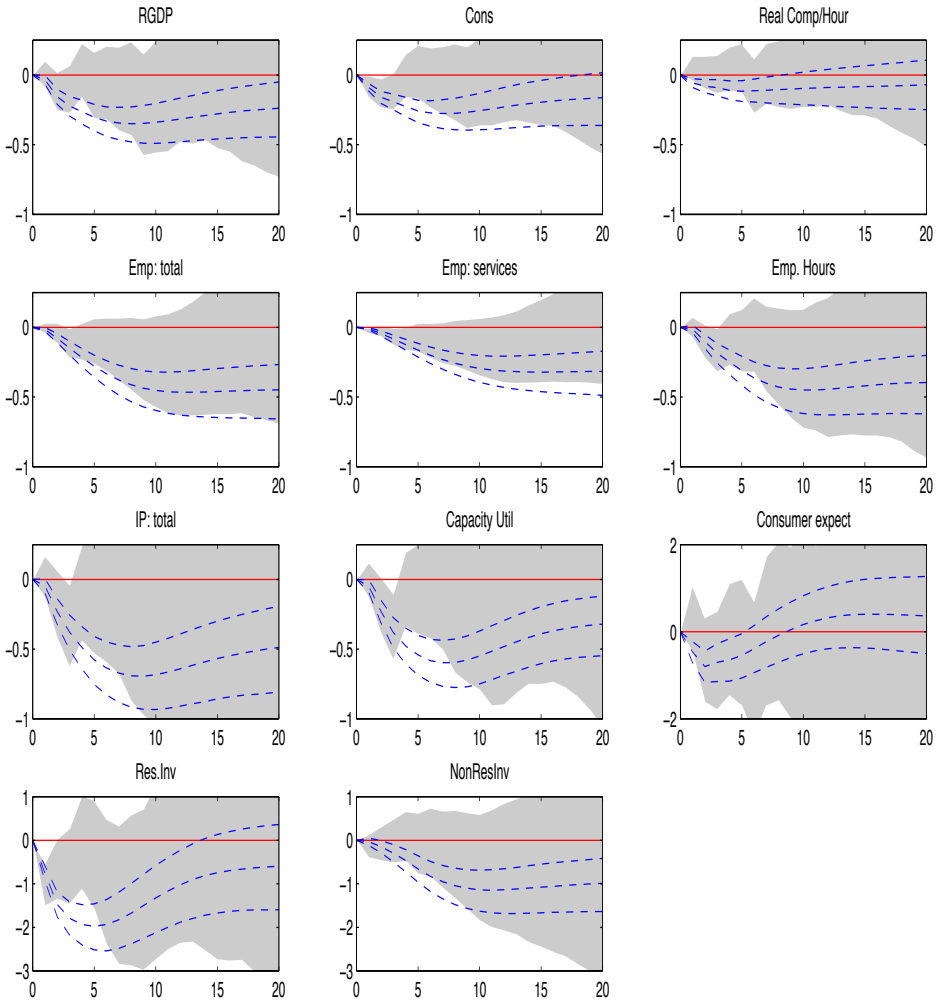
# Figures

Figure 1: Posterior distribution of the hyperparameter governing the standard deviation of the Minnesota Prior



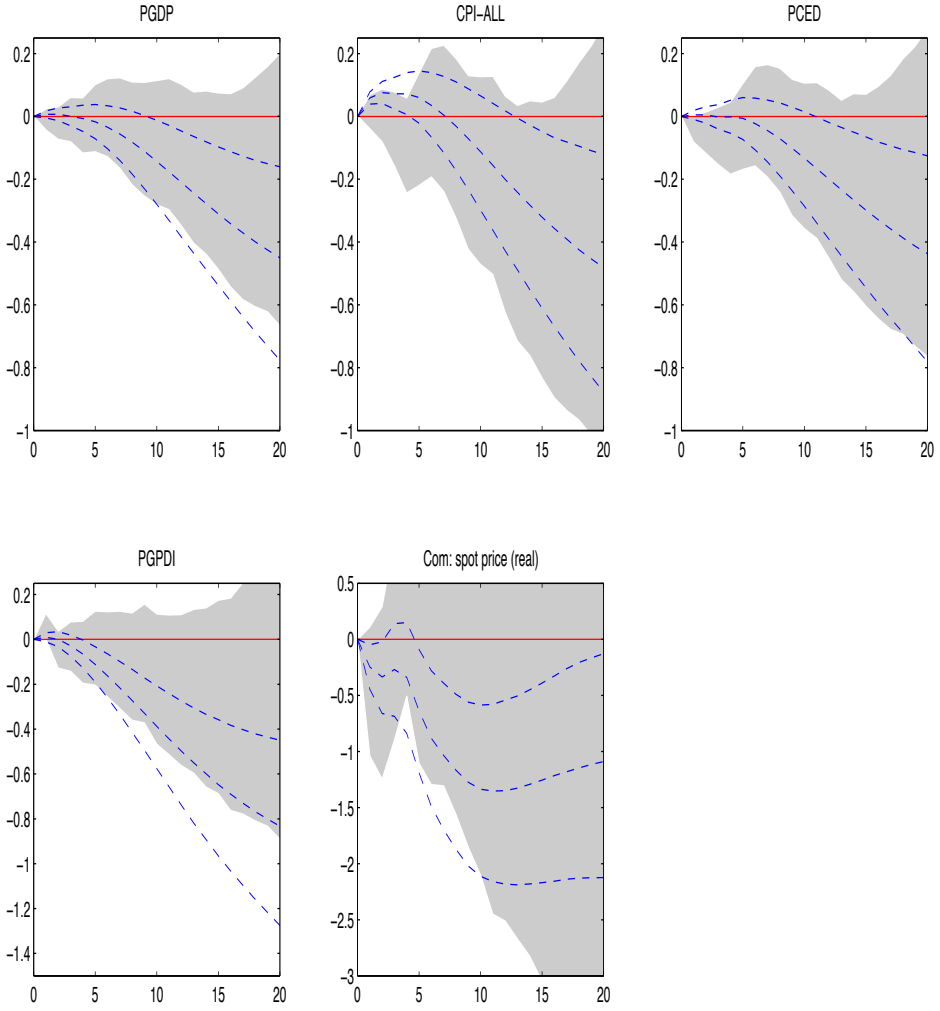

**Note**: The figure reports the posterior distribution of the hyperparameter $\lambda$, the parameter governing the standard deviation of the Minnesota prior in the small, medium, large BVARs, and its prior distribution. The posterior distribution is obtained using the whole sample.

Figure 2: Impulse responses of real variables



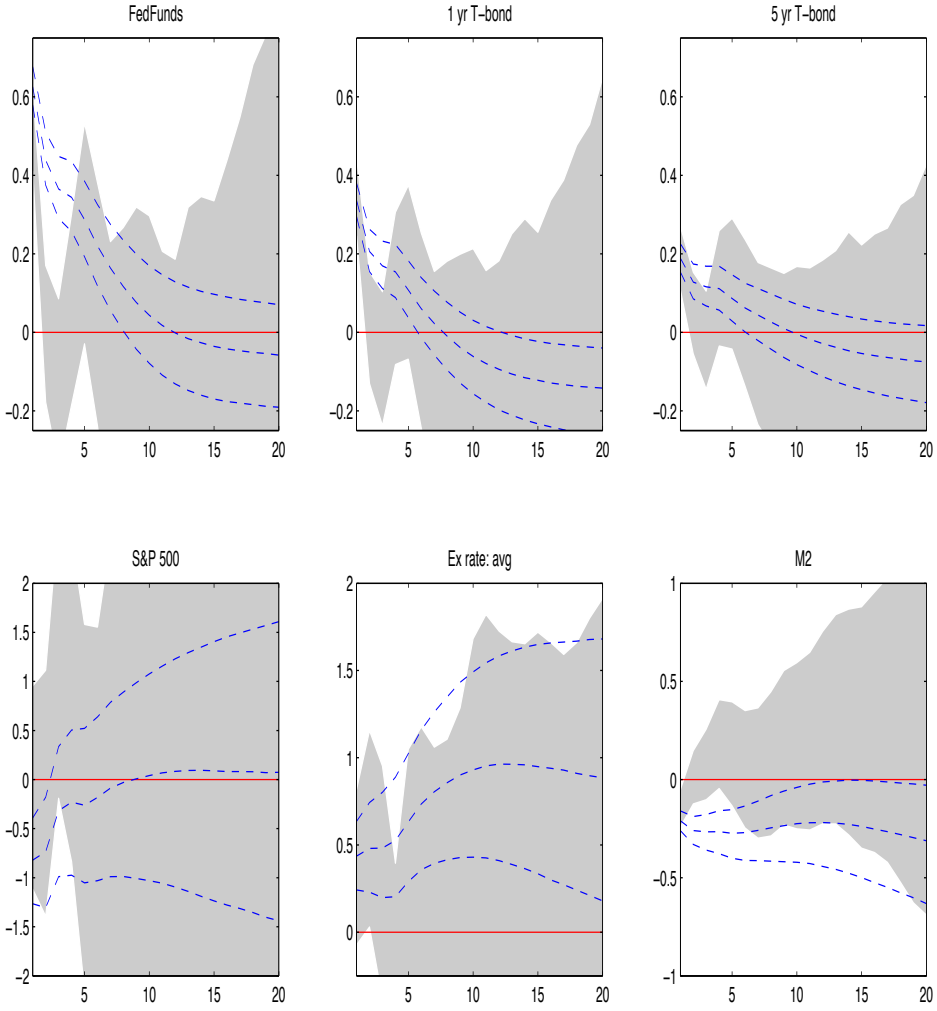

**Note**: The figure reports the 16th, 50th and 84th percentiles (dashed lines) of the distribution of the impulse response functions of the large BVAR to a one standard deviation monetary policy shock. The grey area refers to the corresponding error bands for the flat-prior VAR.

Figure 3: Impulse responses of nominal variables



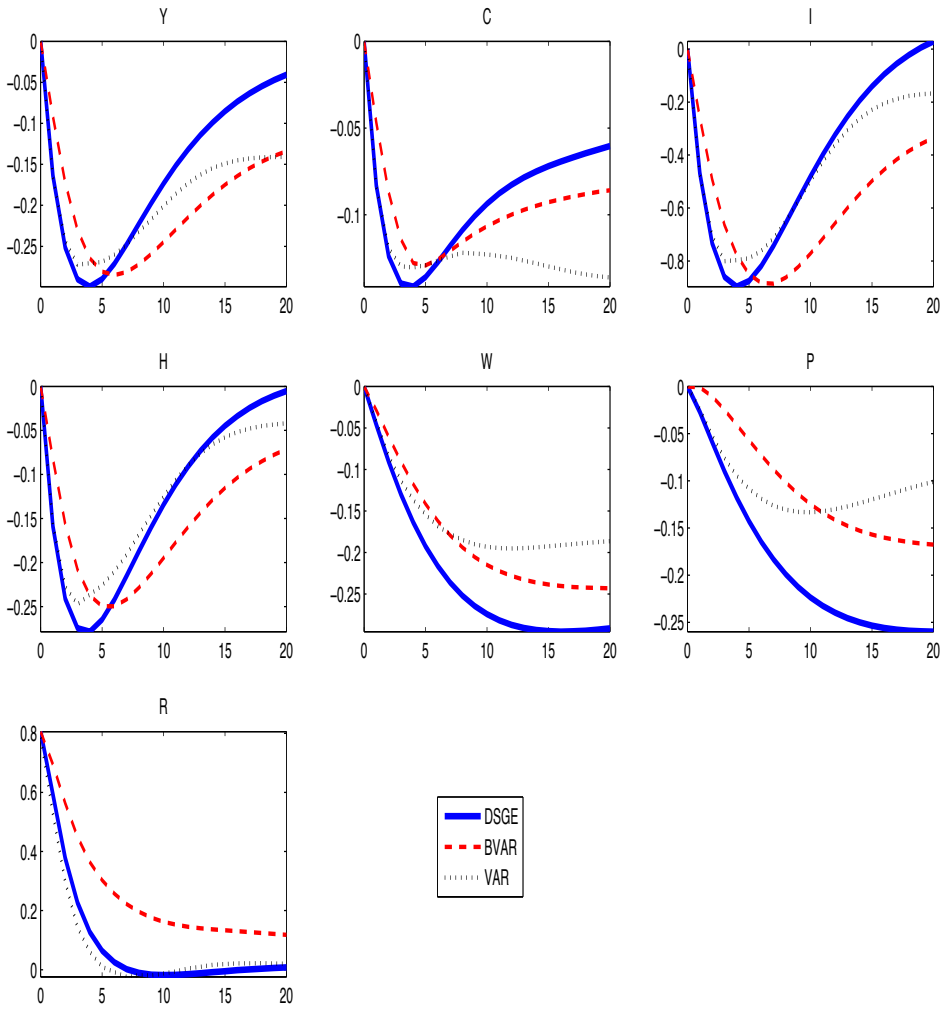

**Note**: The figure reports the 16th, 50th and 84th percentiles (dashed lines) of the distribution of the impulse response functions of the large BVAR to a one standard deviation monetary policy shock. The grey area refers to the corresponding error bands for the flat-prior VAR.

Figure 4: Impulse responses of financial variables



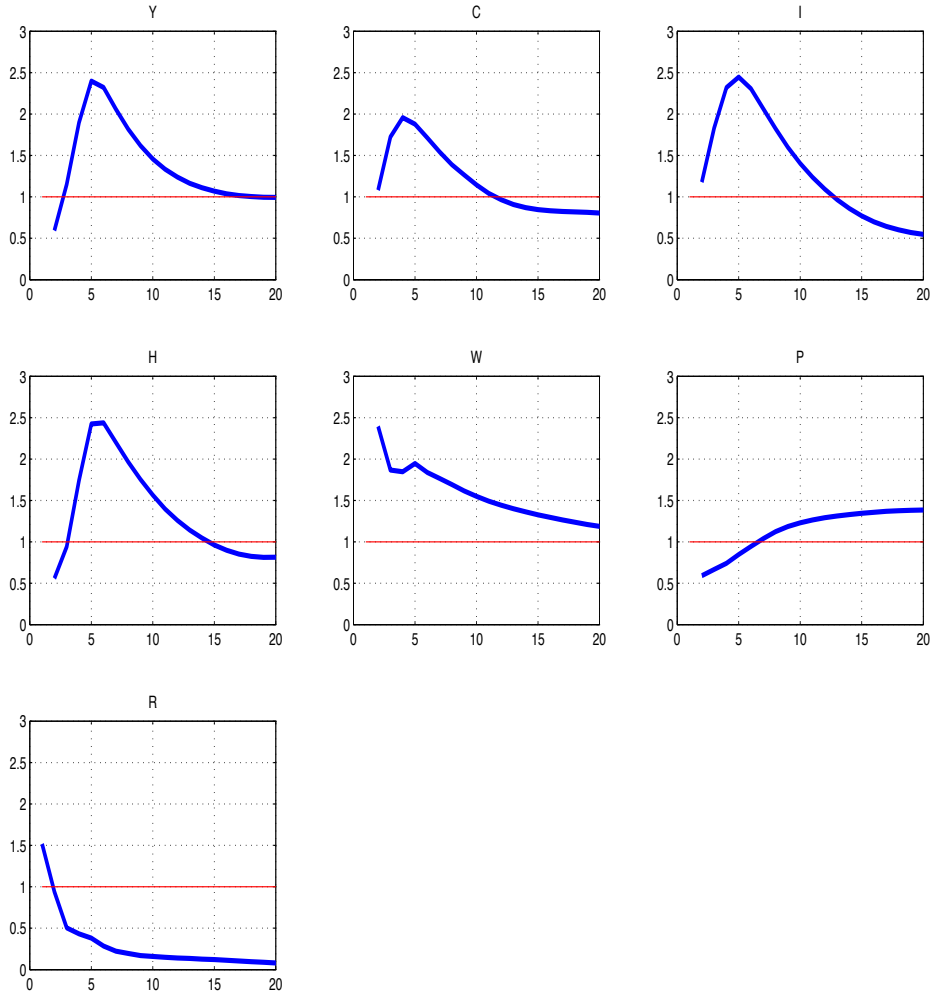

**Note**: The figure reports the 16th, 50th and 84th percentiles (dashed lines) of the distribution of the impulse response functions of the large BVAR to a one standard deviation monetary policy shock. The grey area refers to the corresponding error bands for the flat-prior VAR.

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 Monte Carlo 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.