

# Prior Selection for Vector Autoregressions\*

Domenico Giannone (Université Libre de Bruxelles)  
Michele Lenza (European Central Bank)  
Giorgio Primiceri (Northwestern University)

March 26, 2010

## Abstract

Vector autoregressions (VARs) are flexible time series models that can capture complex dynamic interrelationships among macroeconomic variables. Their generality, however, comes at the cost of being very densely parameterized. As a result, the estimation of VARs tends to deliver good in-sample fit, but unstable inference and inaccurate out-of-sample forecasts, particularly when the model includes many variables. A potential solution to this problem is combining the richly parameterized unrestricted model with parsimonious priors, which help controlling estimation uncertainty. Unfortunately, however, the issue of how to optimally set the weight of the prior relative to the likelihood information is largely unexplored. In this paper, we propose a simple and theoretically founded methodology for prior selection in Bayesian VARs. Our recommendation is to select priors using the marginal data density (i.e. the likelihood function integrated over the model parameters), which only depends on the hyper-parameters that characterize the relative weight of the prior model and the information in the data. We show that the out-of-sample forecasting accuracy of our model not only is superior to that of VARs with flat priors, but is also comparable to that of factor models.

## 1 Prior Information and Forecasting

Throughout most of the paper, we will focus on the following VAR model:

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

where  $y_t$  is an  $n \times 1$  vector of endogenous variables,  $\varepsilon_t$  is an  $n \times 1$  vector of exogenous shocks, and  $C, B_1, \dots, B_p, \Sigma$  are matrices of suitable dimensions containing the model's unknown parameters. With uninformative priors and conditioning on the initial  $p$  observations, the posterior of  $\beta = \text{vec}([C, B_1, \dots, B_p]')$  is centered at the OLS estimate of the coefficients and it is easy to compute. It is well known, however, that working with flat priors yields poor inference, particularly in large dimensional systems. One

---

\*PRELIMINARY

typical symptom of this problem is the poor out-of-sample forecasting performance of these models, 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 standard 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 conjugate prior distribution for the VAR coefficients

$$\beta|\Sigma \sim N\left(b, \Sigma \otimes \Omega \frac{1}{\xi}\right),$$

where  $b$  and  $\Omega$  are given, and  $\xi$  is a scalar controlling the tightness of the prior information. The conditional posterior of  $\beta$  can be obtained by multiplying this prior by the likelihood function and takes the form

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

where  $y \equiv y^T$  is the  $T \times n$  matrix of observed data up to time  $T$ ,  $x = [x_1, \dots, x_T]'$  and  $x_t = [1, y'_{t-1}, \dots, y'_{t-p}]'$ . Notice that an increase of  $\xi$  pulls the posterior mean of  $\beta$  towards the prior mean, while reducing the posterior variance.

One natural way to judge the impact of different priors on the model's forecasting performance is to look at their effect on the probability of observing low forecast errors. To this end, rewrite (1.1) as

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

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

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

which makes it easy to see that neither very high nor very low values of  $\xi$  are likely to be ideal. On the one hand, extremely high values of  $\xi$  generate very concentrated density forecasts, centered around  $X_Tb$ . This results in a low probability of observing small forecast errors, unless the prior mean lies in a close neighborhood of the likelihood peak (which there is no reason to believe). On the other hand, when  $\xi$  is too low and the prior too uninformative, the model generates very dispersed density forecasts (especially in high-dimensional VARs). 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.

Given that neither flat nor dogmatic priors maximize the fit of the model, the literature has proposed a variety of methodologies to optimally 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 overfitting by choosing the shrinkage parameters that yield a desired in-sample fit.<sup>1</sup>

From a purely Bayesian perspective, however, the choice of the informativeness of the prior distribution is conceptually identical to the inference on the “standard” model’s parameters. Suppose, for instance, that a model is described by a likelihood function  $p(y|\theta)$  and a prior distribution  $p(\theta|\gamma)$ , where  $\theta$  denotes the model’s parameters and  $\gamma$  corresponds to the hyperparameters, i.e. those coefficients that do not directly enter the likelihood function but only affect the prior distribution.<sup>2</sup> As usual in Bayesian inference, the estimation of the model’s hyperparameters involves evaluating their posterior distribution, which is given by

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

where  $p(\gamma)$  denotes the prior density on the hyperparameters, while

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

is known as the marginal data density (or marginal likelihood, ML). In other words, the posterior distribution of the hyperparameters is proportional to the product of the ML and the prior density. If such a prior is flat, the shape of the posterior of the hyperparameters is determined only by the ML.

Crucially, in the case of VARs with conjugate priors, the ML is available in closed form. As a consequence, depending on the application, we can easily simulate the posterior of the hyperparameters using Monte Carlo methods, or simply maximizing this posterior and work with the mode (the peak of the ML) and the posterior variance (approximated by the inverse Hessian at the peak).

Observe that estimating hyperparameters using the ML, not only is a theoretically clean way of choosing the informativeness of the prior distribution, but has also several appealing interpretations. First, the ML is a measure the out-of-sample forecasting performance of a model (see Geweke, 2001; Geweke and Whiteman, 2006).

More precisely, the ML corresponds to the probability density that the model generates zero forecast errors. This can be easily seen by rewriting the ML as a product of conditional densities, i.e.

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

Therefore, maximizing the ML corresponds to picking the value of the hyperparameter that maximizes the one-step-ahead out-of-sample forecasting ability of the model.

---

<sup>1</sup>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, 2010).

<sup>2</sup>The distinction between parameters and hyperparameters is mostly fictitious and made only for convenience.

Second, the strategy of estimating hyperparameters by maximizing the ML 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) can be thought of as conducting Bayesian inference on the population parameters of a random effects model or, more generally, a hierarchical model (see, for instance, Gelman, Carlin, Stern, and Rubin, 2004).

Some of these ideas have also been used at times in the applied macroeconomic literature. For example, the ML is commonly used to perform Bayesian model comparison (see Smets and Wouters, 2007, for a recent influential application) or Bayesian model averaging (Sala-I-Martin, Doppelhofer, and Miller, 2004; Wright, 2009).<sup>3</sup>

More directly related to our paper, Del Negro and Schorfheide (2004) and Del Negro, Schorfheide, Smets, and Wouters (2007) use the ML to choose the effective sample size of a prior for VARs derived from the posterior density of a New Keynesian dynamic stochastic general equilibrium model. Primiceri (2005) adopts a similar strategy to set the tightness of the prior on the time variation of parameters and volatilities of a time-varying VAR. In the forecasting literature, Carriero, Kapetanios, and Marcellino (2010) also use the ML of a BVAR to determine the shrinkage of bond yield dynamics towards univariate AR models. We generalize this approach to the optimal selection of a variety of commonly adopted prior distributions for BVARs (see Sims and Zha, 1998; Robertson and Tallman, 1999, for an overview). 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 BVARs. In addition, 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) and empirical Bayes regressions with large sets of predictors (Knox, Stock, and Watson, 2000). However, while Knox, Stock, and Watson (2000) concentrate on single-equation methods, our focus is on multivariate models.

In the next sections of the paper we will illustrate the details of the empirical application of our proposed methodology and evaluate its performance in terms of macroeconomic forecasting.

## 2 The Marginal Likelihood of BVARs

For the implementation of our proposed methodology, we follow the literature and assume that the prior on the unknown parameters of the VAR  $(\beta, \Sigma)$  has a Normal-Inverse-Wishart distribution:

$$\Sigma \sim IW(\Psi; d) \tag{2.2}$$

$$\beta|\Sigma \sim N(b, \Sigma \otimes \Omega). \tag{2.3}$$

---

<sup>3</sup>Geweke and Amisano (2009) use a different, but related approach to model averaging, based on the log-predictive score.

The  $n \times n$  symmetric matrix  $\Psi$  and the scalar  $d$  are the scale matrix and the degrees of freedom of the Inverse-Wishart distribution. We set  $d = n + 2$ , which is the minimum number of degrees of freedom that guarantees that the prior is proper. 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.

As for the conditional Gaussian prior for  $\beta$ , we combine the three most popular prior densities used by the existing literature:

1. The first prior is a generalization of the so-called Minnesota (MN) prior, first introduced by Litterman (1980). The details of the prior specification are reported in Appendix A. Here it suffices to say that the MN prior on the VAR coefficients is centered on the assumption that each variable follows a random walk process. The variance of this prior declines with the lag of the variable multiplying the coefficient. The overall tightness of the MN prior is controlled by the hyperparameter  $\lambda$ .
2. The second prior follows Doan, Litterman, and Sims (1984) and imposes discipline on the sum of coefficients (SOC) in each equation. More specifically, the SOC prior postulates that the sum of coefficients on own lags is centered at 1, while the sum of coefficients on other variables' lags is centered at 0. The hyperparameter  $\mu$  controls the tightness of the SOC prior.
3. The third prior is the so-called single-unit-root prior (SUR), suggested by Sims and Zha (1998), and is motivated by the fact that VAR inference typically conditions on the initial observations. The hyperparameter  $\delta$  controls the tightness of the SUR prior.

These three conditional Gaussian priors for  $\beta$  can be combined and cast into the form of (2.3), where the matrix  $\Omega$  becomes a function of the vector  $(\lambda, \mu, \delta, \psi)$  which controls the shrinkage.

In addition, under the prior (2.2)-(2.3), the ML is available in closed form and, as shown in appendix B, it is given by the following expression:

$$\begin{aligned}
p(y|b, \Omega, \Psi, d) &= \left(\frac{1}{\pi}\right)^{\frac{nT}{2}} \frac{\Gamma_n\left(\frac{T+d}{2}\right)}{\Gamma_n\left(\frac{d}{2}\right)} \\
&\quad |\Omega|^{-\frac{n}{2}} \cdot |\Psi|^{\frac{d}{2}} \cdot |x'x + \Omega^{-1}|^{-\frac{n}{2}} \cdot \\
&\quad \left| \Psi + \hat{\varepsilon}'\hat{\varepsilon} + (\hat{B} - \hat{b})' \Omega^{-1} (\hat{B} - \hat{b}) \right|^{-\frac{T+d}{2}}, \quad (2.4)
\end{aligned}$$

where  $\Gamma_n(\cdot)$  is the  $n$ -variate Gamma function,  $\hat{\varepsilon}$  is the  $T \times n$  matrix of the VAR residuals computed at the posterior mode of the VAR parameters (i.e.  $\hat{B}$ ), and  $\hat{b}$  is a  $(1 + np) \times n$  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. As shown in the previous section, with a flat prior on  $(b, \Omega, \Psi, d)$ , the posterior of the hyperparameters

will be proportional to (2.4). In particular, we set the hyperparameters  $\lambda$ ,  $\mu$ ,  $\delta$  and  $\psi$  to maximize the marginal likelihood, i.e. at their posterior modes.<sup>4</sup>

We now turn to the application of our method to macroeconomic forecasting.

### 3 Forecasting Evaluation of BVAR Models

The assessment of the forecasting performance of econometric models has become standard in macroeconomics, even if the ultimate goal is not out-of-sample prediction. This is because forecasting evaluation can be seen as a validation procedure, which is particularly important for very flexible and general models. In general, introducing complexity in the model to better describe the data does not necessarily enhance real-time forecasting performances. 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 miss-specification. Out-of-sample forecasting reflects both parameter uncertainty and model mis-specification and reveals whether the benefits from more flexibility are or not 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 described by Stock and Watson (2008). The complete database includes 149 quarterly variables in the sample range 1959Q1 - 2008Q4. In this section, we consider a subset of this database that includes the most relevant macroeconomic aggregates.<sup>5</sup> More in details, we estimate three models using progressively larger cross-sections of variables:<sup>6</sup>

1. a *SMALL*-scale model -the prototypical monetary VAR- using three variables, i.e. GDP, GDP deflator and the Federal Funds rate;
2. a *MEDIUM*-scale model, using the variables included in the Smets and Wouters (2007) DSGE model of the US economy. In other words we add consumption, investment, hours worked and wages to the *SMALL* model;
3. a *LARGE*-scale model, with 23 variables, using a dataset that nests the previous two specifications and also includes a number of important additional labor market, financial and monetary variables.

The variables enter the models in annualized (log)-levels<sup>7</sup> and we set the number

---

<sup>4</sup>An alternative that we are exploring is simulating the whole posterior distribution using MCMC methods; this is likely to deliver more accurate results, but it is more time-consuming.

<sup>5</sup>Details about the database can be found in Appendix C. Since several variables are monthly, we follow Stock and Watson (2008) and transform them into quarterly by taking averages.

<sup>6</sup>We are currently experimenting with VAR models including the whole cross-section in Stock and Watson (2008).

<sup>7</sup>In other words, we take logs and multiply by 400 variables like GDP, consumption etc while we do not transform variables that are defined as annualized rates, as interest rates.

of lags to five. We produce the BVAR forecasts recursively. We start with the estimation sample from 1959Q1 to 1969Q4. First, we maximize the marginal likelihood on the estimation sample in order to select the hyperparameters and thus specify the prior distribution of the parameters. We then compute the posterior distribution of these parameters and use the median to iteratively generate forecasts at horizons 1-8 quarters. Finally, we update the sample by one quarter and repeat the procedure until the end of the sample.

As a measure of forecast accuracy, we use point mean squared forecasting errors (MSFE) between the forecasts and a target defined in terms of the  $h$ -period annualized average growth rates:

$$z_{i,t+h}^h = \frac{1}{h} [y_{i,t+h} - y_{i,t}]$$

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.

Table 1 reports the Mean Squared Forecast error of real GDP, the GDP deflator and the Federal Funds Rate obtained with our three BVAR models and their OLS (flat prior) counterparts.

INSERT TABLE 1 HERE

A few results stand out. The forecasts obtained with a BVAR with prior selected by maximizing the marginal likelihood are systematically more accurate than those produced when using flat priors (OLS), for all the three variables at all horizons. When estimated using a flat prior, the forecasts deteriorate substantially when moving from the Small scale to the Medium scale model. This indicates that the gains from exploiting larger information are completely off-set by the increased estimation error. For the large scale model, the flat prior produces predictions that are extremely unreliable especially at the longer horizons of one and two years, reflecting the instability due to the large estimation uncertainty. When estimated with informative priors, the three models produce forecasts that are systematically more accurate than those obtained with flat priors (OLS). Predictions are reliable not only for the small and the medium scale model but also for the large scale model. When increasing the scale of the model, forecast accuracy does not deteriorate, on the contrary accuracy generally improves. In this sense the imposition of prior turns the curse into the blessing of dimensionality. Good performance is already obtained, however, with the medium-size model. Results are in line with Bańbura, Giannone, and Reichlin (2010) who set the prior using a heuristic procedure. The advantage of our methodology is that it is theoretically transparent.

## 4 Single equation forecasting

In the previous section, we have shown that VARs with informative priors perform much better than their flat prior counterparts when we select hyperparameters based on the

marginal likelihood. In this section we instead compare those BVAR forecasts with alternative methods specifically designed to extract information from a large cross-section of predictors. These methods are well known to produce very accurate macroeconomic forecasts.

#### 4.1 Factor augmented regression

Factors offer a parsimonious representation for macroeconomic variables while retaining the salient features of the data that notoriously strongly comove. For this reason, factor augmented regressions are widely used in order to deal with the curse of dimensionality as a large set of potential predictors can be replaced in the regressions by a much smaller number of factors.

In this section we focus on the factor based forecasting approach of Stock and Watson (2002a,b). With respect to the exercises in the previous section, there are two main differences. First, concerning data transformations, factor models are estimated in first differences in order to achieve stationarity. On the contrary, the VAR models of the previous section were specified in log-levels. We denote the stationarized version by  $z_{i,t} = \Delta y_{i,t}$ .<sup>8</sup>

Second, with factor augmented regression it is common practice to use  $h$ -steps ahead projection to construct the multistep forecasts directly.<sup>9</sup> This is different from the previous section, where we produced iterative forecasts.

More in details, consider the following forecasting equation:

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

As in the previous section,  $z_{i,t+h}^h$  denotes the  $h$ -steps ahead variable to be forecasted. The predictors  $f_{k,t}, k = 1, \dots, r$  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. First, the common factors  $f_{k,t}$  are estimated by principal components extracted from the dataset containing all the variables in the large dataset ( $y_t = (y_{1,t}, \dots, y_{n,t})'$ , with  $n = 23$ )<sup>10</sup>. The data are standardized before extracting the factors since principal components are not scale invariant. Second, the coefficients are estimated by ordinary least squares. Using all the principal components (i.e. by setting  $r$  equal to the number of variables  $n$ ) would be equivalent to an OLS estimating of the regression equation on all the available variables.

---

<sup>8</sup>All the qualitative results are confirmed when we difference twice the price variables as in Stock and Watson (2002a,b).

<sup>9</sup>Iterated forecasts using factor models have been performed by specifying a BVAR on the common factors and using Kalman filtering techniques, see Giannone, Reichlin, and Small (2008).

<sup>10</sup>Qualitative results are confirmed when using the whole cross-section in Stock and Watson (2008)

The parameters of the forecasting equations are set as in Stock and Watson (2008):  $p_z = 4$  and  $r = 3$ . For convenience we also report the results for the autoregressive model which corresponds to  $r = 0$ .

## 4.2 Univariate Bayesian Regression

In this sub-section we generate forecasts regressing the dependent variable directly on the large cross-section of predictors. If the latter is large, classical estimation techniques cannot handle this problem. Bayesian shrinkage deals instead with the curse of dimensionality and can be seen as an alternative to factor augmented regressions. This idea has been recently put forward by De Mol, Giannone, and Reichlin (2008), who also provide theoretical rates for the specification of the degree of shrinkage in the prior distributions. In particular, it is shown that the degree of shrinkage should be related asymptotically to the number of predictors, giving some guidance in the specification of the priors. The marginal likelihood, however, provides a clean and objective method to choose the prior tightness also in finite samples.

More in details, our forecasting model is:

$$z_{i,t+h} = c_i + \sum_{s=0}^{p_z-1} \alpha_{i,s} z_{i,t-s} + \sum_{s=0}^{p_z-1} \sum_{j \neq i} \beta_{ij,s} y_{j,t-s} + e_{i,t+h}^h$$

The errors are assumed to be independently, identically and normally distributed with mean zero and variance  $\sigma^2$ .<sup>11</sup> We impose the prior that all the coefficients are independently and normally distributed with mean zero and variance  $E[(\alpha_{i,s})^2] = \frac{\xi_1 \sigma^2}{s^2}$  and  $E[(\beta_{ij,s})^2] = \psi_i^2 \frac{\xi_2 \sigma^2}{s^2}$ .

The parameter  $\xi_1$  controls for the degree of shrinkage for the lags of the target variable. The parameter  $\xi_2$  controls the degree of shrinkage for all the other predictors. Longer lags are shrunk more, along the lines of the Minnesota prior. The parameter  $\psi_i^2$  is specific to each predictor, controls for the different scale of the variables and it is set equal to the sample variance of each predictors  $z_{i,t}$ .<sup>12</sup>

We set  $p_z = 4$ . The parameters  $\xi_1$ ,  $\xi_2$  and  $\sigma^2$  are selected by maximizing the marginal likelihood.

Forecasting results using principal components and for the univariate Bayesian regression are reported, respectively, in the sixth and seventh columns of Table 2. The Table report Mean Squared Forecasting Errors relative to the autoregressive model. For comparison we also report the results for the BVAR model. We focus on one year ahead predictions; qualitative results are confirmed at other horizons. Results indicate that the BVAR predictions are competitive with those produced using principal components. The performances of the univariate regressions are comparable to those of factor augmented regressions. The forecasts are also highly correlated. This is in line

<sup>11</sup>Remark: the independence of the errors does not hold for  $h > 1$ . The model in this case is clearly misspecified. We will not deal with this for the moment to be as close as possible with traditional literature that estimate the forecasting equation by Ordinary Least Squares.

<sup>12</sup>Notice that this is equivalent to running the regression on standardized variables and using an homogenous degree of shrinkage across variables.

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, our estimates of the factors, are strictly connected.

## 5 Conclusion

TO BE ADDED

## References

- BAÑBURA, M., D. GIANNONE, AND L. REICHLIN (2010): “Large Bayesian VARs,” *Journal of Applied Econometrics*, 25(1), 71–92.
- 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.
- CARRIERO, A., G. KAPETANIOS, AND M. MARCELLINO (2009): “Forecasting exchange rates with a large Bayesian VAR,” *International Journal of Forecasting*, 25(2), 400–417.
- CARRIERO, A., G. KAPETANIOS, AND M. MARCELLINO (2010): “Forecasting Government Bond Yields,” mimeo, University of London.
- 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 G. AMISANO (2009): “Optimal Prediction Pools,” Working Paper Series 1017, European Central Bank.
- 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,” Discussion paper.
- 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.
- GIANNONE, D., L. REICHLIN, AND D. SMALL (2008): “Nowcasting: The real-time informational content of macroeconomic data,” *Journal of Monetary Economics*, 55(4), 665–676.
- 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. (2010): “Forecasting with Medium and Large Bayesian VARs,” Manuscript, University of Strathclyde.
- LITTERMAN, R. B. (1980): “A Bayesian Procedure for Forecasting with Vector Autoregression.,” Working paper, Massachusetts Institute of Technology, Department of Economics.
- 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.
- 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.
- SALA-I-MARTIN, X., G. DOPPELHOFER, AND R. I. MILLER (2004): “Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach,” *American Economic Review*, 94(4), 813–835.
- 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.
- 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.
- WRIGHT, J. H. (2009): “Forecasting US inflation by Bayesian model averaging,” *Journal of Forecasting*, 28(2), 131–144.

## Tables

Table 1: BVAR MSFE

Steps Ahead		Small (S)		Medium (M)		Large (L)		Extra Large (XL)	
		OLS	BVAR	OLS	BVAR	OLS	BVAR	OLS	BVAR
One Quarter	Real GDP	13.07	10.46	23.03	8.92	77.32	8.79		
	GDP Deflator	2.33	1.50	3.65	1.42	15.14	1.35		
	Federal Funds Rates	1.67	1.14	2.25	1.13	6.62	1.04		
One Year	Real GDP	5.15	4.16	17.35	3.74	152.50	3.44		
	GDP Deflator	2.28	1.53	4.94	1.34	54.48	1.37		
	Federal Funds Rates	0.63	0.41	0.94	0.33	64.34	0.33		
Two Years	Real GDP	5.41	2.70	18.36	3.02	75152.00	2.67		
	GDP Deflator	3.68	2.61	7.59	1.96	18351.00	2.46		
	Federal Funds Rates	0.32	0.21	0.58	0.15	42409.00	0.18		

Table 2: Mean Square Forecasting Error relative to the Autoregressive Model

	BVAR S	BVAR M	BVAR L	BVAR XL	PC S&W	Univariate Bayesian Regression
RGDP	0.84	0.75	0.69		0.82	0.80
PGDP	0.86	0.75	0.76		0.92	0.75
FedFunds	0.92	0.73	0.74		0.94	0.91
Cons		0.84	0.69		0.84	0.72
Emp.Hours		0.72	0.72		0.76	0.82
RealComp/Hour		1.01	0.91		1.04	0.93
GPDInv		0.77			0.71	0.86
Res.Inv			0.58		0.72	0.85
NonResInv			0.84		0.88	0.97
CPI-ALL			0.95		0.89	0.77
Com:spotprice(real)			0.88		1.00	0.97
IP:total			0.59		0.70	0.80
Emp:total			0.67		0.83	0.81
Emp:services			0.76		0.80	0.84
Cons			0.67		0.84	0.72
PCED			0.98		0.99	0.84
PGPDI			0.60		0.82	0.72
CapacityUtil			0.66		0.75	0.82
Consumerexpect			0.90		0.98	0.89
Emp.Hours			0.70		0.75	0.81
RealComp/Hour			0.91		1.04	0.93
1yrT-bond			0.81		1.01	0.98
5yrT-bond			0.93		1.07	0.99
S&P500			1.11		1.15	1.09
Exrate:avg			1.00		1.07	1.13
Reservestot			0.99		0.99	0.99
M2			0.89		0.97	0.83

## Appendix A: Details on the prior distributions

This appendix describes the details of the three conditional Gaussian priors on  $\beta$  that we adopt.

1. Our first prior is a generalization of the so-called Minnesota (MN) prior, first introduced by Litterman (1980). The MN prior on the VAR coefficients is centered on the assumption that each variable follows a random walk process. More precisely, this prior states that

$$(B_s)_{ij} | \Sigma \sim N \left( \phi_{ijs}; \frac{1}{\lambda^2} \frac{1}{s^2} \frac{\Sigma_{ii}}{\psi_j / (d - n - 1)} \right),$$

with

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

In particular, the variance of the prior depends on whether a parameter multiplies a distant lag of a variable. Moreover, coefficients multiplying the same variable and lag in different equations are allowed to be correlated. Finally, the hyperparameter  $\lambda$  controls the overall tightness of the prior. There are no general rules available in the literature to set  $\lambda$ . In our empirical application, we will choose  $\lambda$  using the ML-based approach described above.

2. Our second prior follows Doan, Litterman, and Sims (1984) and imposes discipline on the sum of coefficients in each equation. More specifically, the SOC prior postulates that

$$\sum_{s=1}^p (B_s)_{ij} \sim N \left( f_{ij}, \frac{1}{\mu^2} \frac{\Sigma_{ii}}{\bar{y}_i} \right),$$

where

$$f_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

In this formulation,  $\bar{y}_i$  is the  $i$ th element of the mean of the first  $p$  observations. The hyperparameter  $\mu$  controls the overall tightness of the SOC prior and will be set based on the ML.

3. Finally, we also adopt the so-called single-unit-root prior (SUR), suggested by Sims and Zha (1998). Such a prior is motivated by the fact that inference in VARs ignores the effect of the initial observations (see ???). This prior states that

$$\left[ \bar{y} \left( \sum_{s=1}^p B_s - I \right) + C \right] \sim N \left( 0, \frac{\Sigma}{\delta^2} \right),$$

where the hyperparameter  $\delta$  controls the tightness.

## Appendix B: 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 (??)

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

and rewrite it as

$$\begin{aligned} Y &= X\beta + \varepsilon \\ \varepsilon &\sim N(0, \Sigma \otimes I_T), \end{aligned}$$

where  $y \equiv [y_1, \dots, y_T]'$ ,  $Y \equiv \text{vec}(y)$ ,  $x_t = [1, y'_{t-1}, \dots, y'_{t-p}]'$ ,  $x = [x_1, \dots, x_T]'$ ,  $X = I_n \otimes x$ ,  $\varepsilon \equiv [\varepsilon_1, \dots, \varepsilon_T]'$ ,  $\varepsilon \equiv \text{vec}(\varepsilon)$ ,  $B = [C, B_1, \dots, B_p]'$  and  $\beta = \text{vec}([C, B_1, \dots, B_p]')$ . Finally, define the number of regressors for each equation by  $k \equiv np + 1$ .

As in section (??), the prior on  $(\beta, \Sigma)$  is given by the following Normal-Inverse-Wishart distribution<sup>13</sup>

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

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

The un-normalized posterior of  $(\beta, \Sigma)$  can be obtained by multiplying the prior density by the likelihood function

$$\begin{aligned} p(\beta, \Sigma|Y) &= \left(\frac{1}{2\pi}\right)^{\frac{n(T+k)}{2}} |\Sigma|^{-\frac{T+k+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}\text{tr}(\Psi\Sigma^{-1})}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \\ &\quad e^{-\frac{1}{2} \left[ \begin{aligned} &(Y - X\beta)' (\Sigma \otimes I_T)^{-1} (Y - X\beta) + \\ &+ (\beta - b)' (\Sigma \otimes \Omega)^{-1} (\beta - b) \end{aligned} \right]}. \end{aligned} \quad (5.5)$$

Tedious algebraic manipulations of (5.5) yield the expression

$$\begin{aligned} p(\beta, \Sigma|Y) &= \left(\frac{1}{2\pi}\right)^{\frac{n(T+k)}{2}} |\Sigma|^{-\frac{T+k+n+d+1}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{e^{-\frac{1}{2}\text{tr}(\Psi\Sigma^{-1})}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} \\ &\quad e^{-\frac{1}{2} \left[ \begin{aligned} &(\beta - \hat{\beta})' \left[ X' (\Sigma \otimes I_T)^{-1} X + (\Sigma \otimes \Omega)^{-1} \right] (\beta - \hat{\beta}) + \\ &+ (\hat{\beta} - b)' (\Sigma \otimes \Omega)^{-1} (\hat{\beta} - b) + \hat{\varepsilon}' (\Sigma \otimes I_T)^{-1} \hat{\varepsilon} \end{aligned} \right]}, \end{aligned} \quad (5.6)$$

<sup>13</sup>We are using the following parameterization of the Inverse Wishart density:  $p(\Sigma|\Psi, d) = \frac{|\Psi|^{\frac{d}{2}} \cdot |\Sigma|^{-\frac{n+d+1}{2}} \cdot e^{-\frac{1}{2}\text{tr}(\Psi\Sigma^{-1})}}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)}$ .

where  $\hat{B} \equiv (x'x + \Omega^{-1})^{-1} (x'y + \Omega^{-1}b)$ ,  $\hat{\beta} \equiv \text{vec}(\hat{B})$ ,  $\hat{\varepsilon} \equiv y - x\hat{B}$  and  $\hat{\varepsilon} \equiv \text{vec}(\hat{\varepsilon})$ . It can be shown that (5.6) is the kernel of the following Normal-Inverse-Wishart posterior distribution:

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

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:

$$p(Y) = \int \int p(Y|\beta, \Sigma) \cdot p(\beta|\Sigma) \cdot p(\Sigma) d\beta d\Sigma. \quad (5.7)$$

Let's start with the integral with respect to  $\beta$ . Substituting (5.6) into (5.7) we obtain

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

which can be solved by “completing the squares,” yielding

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

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

$$\begin{aligned}p(Y) &= \left(\frac{1}{2\pi}\right)^{\frac{nT}{2}} |\Omega|^{-\frac{n}{2}} |\Psi|^{\frac{d}{2}} \frac{1}{2^{\frac{nd}{2}} \cdot \Gamma_n\left(\frac{d}{2}\right)} |x'x + \Omega^{-1}|^{-\frac{n}{2}} \\ &\quad \int \left[ \begin{aligned} &|\Sigma|^{-\frac{T+n+d+1}{2}} e^{-\frac{1}{2}\text{tr}(\Psi\Sigma^{-1})}. \\ &e^{-\frac{1}{2} \left[ \underbrace{(\hat{\beta} - b)' (\Sigma \otimes \Omega)^{-1} (\hat{\beta} - b) + \hat{\varepsilon}' (\Sigma \otimes I_T)^{-1} \hat{\varepsilon}}_P \right]} \end{aligned} \right] d\Sigma. \quad (5.8)\end{aligned}$$

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

$$\text{vec}(A)' (D \otimes B) \text{vec}(C) = \text{tr}(A'BCD').$$

This yields

$$P = \text{tr} \left[ \hat{\varepsilon}' \hat{\varepsilon} \Sigma^{-1} + (\hat{B} - \hat{b})' \Omega^{-1} (\hat{B} - \hat{b}) \Sigma^{-1} \right]. \quad (5.9)$$

We can now solve the integral by substituting (5.9) into (5.8), 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:

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

## Appendix C: Data

Table 3: The description of the database

Variables	Mnemonic	Transf. BVAR	Transf. Univariate	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