วารสารเศรษฐศาสตร์ธรรมศาสตร์ **Thammasat Economic Journal**
ปีที่ 28 ฉบับที่ 1 มีนาคม 2553 **Vol.28, No.1, March 2010**

Forecasting with Bayesian VARs: Does Larger Mean Better?

Pawin Siriprapanukul[*]

บทคัดย่อ

ในทางทฤษฎีนั้นเราน่าจะเพิ่มความสามารถในการทำนายของแบบจำลอง Bayesian Vector Autoregressions (Bayesian VARs) ได้จากการเพิ่มตัวแปรเข้าไปในแบบจำลองดังกล่าว บทความนี้จะทำการทดสอบสมมุติฐานข้างต้นในเชิงประจักษ์ โดยเราจะเปรียบเทียบความสามารถในการทำนายตัวแปรสำคัญทางเศรษฐกิจ 3 ตัวแปรของแบบจำลอง Bayesian VAR ขนาดใหญ่ที่อาศัยตัวแปรทั้งสิ้น 131 ตัวแปร กับแบบจำลองขนาดเล็กอื่นๆ โดยที่แบบจำลองขนาดเล็กที่สุดจะอาศัยตัวแปรเพียงแค่ 3 ตัวเท่านั้น ในการเปรียบเทียบความสามารถในการทำนายครั้งนี้เราให้ความสำคัญกับค่า hyperparameter ตัวหนึ่ง ซึ่งทำหน้าที่เป็นตัวกำหนดค่าความแปรปรวนโดยรวมของตัว Prior Distribution ในการประมาณค่าแบบ Bayesian ด้วย เราพบว่าการกำหนดค่า hyperparameter ดังกล่าวจะส่งผลกระทบต่อความสามารถในการทำนายของแบบจำลองขนาดต่างๆ เป็นอย่างมาก ภายหลังจากความพยายามหาค่า hyperparameter ที่เหมาะสมให้กับแต่ละแบบจำลองผลลัพธ์ที่ได้จากการศึกษาของเราสนับสนุนแนวคิดที่ว่าแบบจำลอง Bayesian VAR ขนาดใหญ่จะมีความสามารถในการทำนายที่เหนือกว่าแบบจำลองขนาดเล็กกว่า

Abstract

Conceptually, the impressive forecasting performance of the Bayesian VARs may be further improved by expanding the number of variables into the models. This paper compares the

[*] I am grateful to Farshid Vahid for his valuable suggestions and comments. I also would like to thank Paul Frijter and participants at the 2008 ANU-QUT PhD Workshop for useful comments.

forecasting performance of a large 131 variable Bayesian VAR to much smaller models. Our paper gives especially careful consideration to the effect that a hyperparameter governing the overall tightness of the prior distribution can have, since the performance of a Bayesian regression can be so affected by it. Our results support the idea that larger Bayesian VARs perform better than smaller ones. However, when the hyperparameter of the prior distribution of a smaller model is carefully chosen, the improvement in performances of larger models is not as impressive as previously thought. Even a 3-variable model with an appropriately chosen shrinkage parameter will produce much better forecasts than those reported in the literature.

JEL Classification: C11, C33, C52, C53, E37

Keywords: Bayesian VARs, Macroeconomic Forecasting, Model Specification

1. Introduction

In forecasting macroeconomic variables, Bayesian VARs have an excellent record of performing well in the literature. For example, Robertson and Tallman (1999) report that various Bayesian VAR specifications outperform unrestricted VARs, while Litterman (1986) shows that a Bayesian VAR outperforms an ARIMA, a univariate AR, and the best known commercial forecasting services in out-of-sample forecasting.

According to Litterman(1986), there are at least two advantages to using Bayesian VARs over other nonstructural econometric models. First, since there are many relationships among macroeconomic variables not fully understood by economists, and Bayesian VARs, which allow some uncertainty over the true structure of the economy, give better forecasts than other models that are fully based on just a single economic structure. Second, under situations of a limited observations, Bayesian VARs allow the incorporation of more information. A larger number of parameters can be fitted into the model by assigning appropriate weights to the prior information.

With these advantages, one may argue that larger Bayesian VARs can outperform smaller models in forecasting. Since the exact structure of the economy is not known and the problem about the degrees of freedom is ameliorated, larger Bayesian VARs seem to have an advantage over smaller ones.

Recent forecasting literature is also supportive of the practice of incorporating a large number of variables into models. Many methods have been proposed to allow this practice. These include, for example, the dynamic factor models of Stock and Watson (2002a) and Forni, Halli, Lippi, and Reichlin (2000), and the factor-augmented VAR of Bernanke, Boivin, and Eliasz (2005). There is a lot of evidence demonstrating that this practice improves the forecasting performances of the models. See, for example, D'Agostino and Giannone (2007), Bernanke and Boivin (2003), Stock and Watson (2002b), and Forni, Halli, Lippi, and Reichlin (2003).

Banbura, Giannone, and Reichlin (2008) (henceforth BGR) show that the method of Bayesian VAR admits a large number of endogenous variables. They investigate empirically whether this practice is desirable. According to the authors, a large Bayesian VAR with 131 variables performs better than smaller models with 3, 7, and 20 variables in out-of-sample forecasting. The largest model clearly outperforms the two smallest ones, its forecasting performance was however, matched by the model with 20 variables.

The Bayesian VAR estimator, however, depends on a hyperparameter determining the relative weight given to the prior information, and as a consequence the out-of-sample forecasting performance of a model is influenced by this hyperparameter as well. BGR's findings therefore are based on the particular way that they determine the value of this hyperparameter. We do not find the BGR's method the most natural way of setting this value, and there is no reason to believe that their results will be robust if this parameter value is chosen in a different way. In section 3, we show that if we assign different values to this hyperparameter, larger VARs of BGR may not outperform smaller ones.

This paper first determines a suitable hyperparameter value for each model, which makes the most out of each model given our pre-evaluation period. Given a model and a forecast horizon, we find the hyperparameter value that minimizes the magnitude of out-of-sample forecast errors in a part of the pre-evaluation period. After that, we assign this suitable value to that model during our out-of-sample assessment in an evaluation period. This is shown in section 4. Our result in this section supports BGR's finding that larger models perform better in the overall picture. However, the performances of the larger models are not dramatically different from that of the smallest model.

We realize that the suitable hyperparameter value can vary over time. The time-varying hyperparameter may affect different models in different magnitudes. To make our study more robust, we extend it out in two additional ways. First, we allow the suitable hyperparameter value of each model to change every 10 years. For each additional 10 years of observations, we re-calculate a suitable hyperparameter value for each model. After that, we use this hyperparameter value in making forecasts for the next 10 years, at which point we again re-calculate a new hyperparameter value. We assess the performances of Bayesian VARs under this practice. The result of this is shown in section 5. Contrary to our expectation, this practice does not improve the forecasting performances of any model specifications.

Second, we apply an updating scheme for the hyperparameter value. With additional data, we calculate the effect of a small change in the value of the hyperparameter. If the change signals an improvement in the forecasting performance of a model, a new hyperparameter value is applied to the model for making the forecast for the next period. Section 6 reports the result from this experiment. It shows that our updating scheme marginally improves the forecasting performance of each model specification.

Section 2 shows the details of the model and the estimation method used in this paper, and section 7 concludes the paper.

2. Estimated Model

We estimate the same Bayesian VARs as BGR. Let $Y_t = (y_{1,t}\ y_{2,t}\ ...\ y_{m,t})'$ be an $m \times 1$ column vector of *m* endogenous variables in period *t*. The reduced form of the VAR is:

$$\underset{m\times1}{Y_t} = \underset{m\times m}{A_1}\,\underset{m\times1}{Y_{t-1}} + \underset{m\times m}{A_2}\,\underset{m\times1}{Y_{t-2}} + ... + \underset{m\times m}{A_p}\,\underset{m\times1}{Y_{t-p}} + \underset{m\times1}{\mathbf{c}} + \underset{m\times1}{U_t}, \tag{1}$$

where $\mathbf{c} = (c_1, ..., c_m)'$ is the vector of constants, and $U_t = (u_{1,t}\ ...\ u_{m,t})'$ is the vector of unknown disturbances. We assume that:

$$\underset{m\times1}{U_t} \sim N(\underset{m\times1}{\mathbf{0}}, \underset{m\times m}{\Psi}),$$

where the time-invariant matrix Ψ is a positive definite matrix.

Let $X_t = (Y_{t-1}^{'}, ..., Y_{t-p}^{'}, 1)'$ be a column vector containing *p* lags of Y_t and a constant 1. With observations $t = 1, ..., T,$ we can rearrange the VAR from (1) into:

$$\underset{T\times m}{Y} = \underset{T\times k}{X}\,\underset{k\times m}{B} + \underset{T\times m}{U}, \tag{2}$$

where $Y = (Y_1, ..., Y_T)'$ is the matrix of dependent variables, $X = (X_1, ..., X_T)'$ is the matrix of independent variables, $B = (A_1, ..., A_p, c)'$ is the matrix of unknown coefficients, $U = (U_1, ..., U_T)'$ is the matrix of disturbances, and *k* = *mp*+1 is the total number of independent variables. Let **u** be the column vector obtained by stacking the columns of the disturbance matrix *U* from (2). The above assumption on U_t is equivalent to:

$$\underset{Tm\times1}{\mathbf{u}} \sim N(\underset{Tm\times1}{\mathbf{0}}, \underset{m\times m}{\Psi} \otimes \underset{T\times T}{I}),$$

where $\otimes$ represents the Kronecker product, and *I* is an identity matrix.

With the seemingly-unrelated-regressions (SUR) structure, the efficient estimator for B is the same as an unrestricted OLS estimator, which is:

$$\hat{B} = (X'X)^{-1}(X'Y). \tag{3}$$

A major problem with this estimator is that increases in the number of endogenous variables m or the number of lags p used in the model, while hoding the number of observations T finite, render the estimator more unreliable or even uncomputable. Bayesian VARs help avoid this problem.

According to the Bayesian VAR approach, the coefficients in the model are treated as random variables, with given means and variances. The prior information about these means and variances is imposed, and we update this information with the sample observations, using Bayes' law. The end result is the posterior distribution of the coefficients with estimated means and variances. With suitable adjustment to the parameter of the model, there is no requirement on the total number of observations. This is because these observations are only used to update the prior distribution.

The main issue of implementing Bayesian VARs is about the specification of prior distribution. Litterman (1986) suggests imposing a form of prior distributions, generally referred to as a "Minnesota prior". The prior puts the means of the coefficients at the point that makes Y_t be a vector of univariate random walks, i.e. the means are at $A_1 = I_{m\times m}$ and $A_2,\ldots,A_p = \mathbf{0}_{m\times m}$. It may or may not allow for drift. The coefficients are also uncorrelated with each other, with prior variances given by:

$$\text{var}[(A_l)_{ij}] = \begin{cases} \dfrac{\lambda^2}{l^2}, & i = j, \\ \pi\dfrac{\lambda^2}{l^2}\dfrac{\sigma_i^2}{\sigma_j^2}, & \text{otherwise,} \end{cases}$$

where $(A_l)_{ij}$ is the *ij*-th element of the *l*-th lag coefficient matrix A_l, $\lambda \geq 0$ is the hyperparameter determining the overall tightness of the distribution around the random walk, σ_i^2, $i = 1,...,m$, is the variance of disturbance term of the variable $y_{i,t}$ in the VAR, and $\pi \in (0,1]$ is another hyperparameter, reflecting the relative importance of other endogenous variables $j \neq i$ in accounting for the variation of variable *i*. The prior on the intercept **c** is diffuse, i.e. the variance is very high.

Recall that a variance close to zero indicates the distribution is very tight around the mean value. Lowering the value of λ toward zero means tightening the prior distribution toward the random walk. The term l^2 is added to reflect that the longer lagged variables should each have a progressively smaller effect on the current variation of each variable, i.e. that the coefficients in front of these variables should be tightened more toward zero. The hyperparameter π has the same function as l^2, but for other endogenous variables $j \neq i$. It captures the idea that in explaining the variation of a variable, own lags are more important than lags of the other variables. At last, the ratio σ_i^2 / σ_j^2 is used to account for the difference in the units of measurement of different variables *i* and *j*. For more detailed discussion on the prior variances, see Litterman (1986) or Robertson and Tallman (1999).

Early Bayesian VARs assumed the covariance matrix Ψ to be diagonal, fixed, and known. This is considered to be very restrictive. The prior distribution imposed in this model, as recommended by Kadiyala and Karlson (1997), is assumed to be a Normal-(Inverted)-Wishart, which has the form:

$$\underset{km\times 1}{\mathbf{b}} \mid \underset{m\times m}{\Psi} \sim N(\underset{km\times 1}{\tilde{\mathbf{b}}}, \underset{m\times m}{\Psi} \otimes \underset{k\times k}{\tilde{\Omega}}) \quad \text{and} \quad \underset{m\times m}{\Psi} \sim iW(\underset{m\times m}{\tilde{\Psi}}, \alpha), \tag{4}$$

where **b** is the column vector obtained by stacking columns of the matrix *B* from (2). The degree of freedom of the inverted-Wishart distribution is set at $\alpha = m + 2$. This makes the prior mean and variance of the coefficients to be $E(\mathbf{b}) = \tilde{\mathbf{b}}$ and $\text{var}(\mathbf{b}) = \tilde{\Psi} \otimes \tilde{\Omega}$.

Following Kadiyala and Karlson (1997) and BGR, the parameters of the distribution in (4), $\tilde{\mathbf{b}}$, $\tilde{\Omega}$, and $\tilde{\Psi}$, are chosen to match the Minnesota prior. The parameter $\tilde{\mathbf{b}}$ is obtained by stacking columns of the matrix $\tilde{B}$, given by:

$$\underset{k\times m}{\tilde{B}} = \begin{bmatrix} diag(\delta_1,...,\delta_m) \\ \cdots\cdots\cdots\cdots\cdots \\ \underset{(k-m-1)\times m}{\mathbf{0}} \\ \cdots\cdots\cdots\cdots\cdots \\ b_1 \; ... \; b_m \end{bmatrix},$$

where $diag(\delta_1,...,\delta_m)$ is an $m\times m$ diagonal matrix with values $\delta_1,...,\delta_m$ along its main diagonal, δ_i, i = 1,...,m, can be either 0 or 1, and b_i, i = 1,...,m, is a constant or zero. Originally, Litterman sets each δ_i equal to 1. However, following BGR, it is more appropriate to set this value at $\delta_j = 0$ for any mean-reverting variable j.

The parameters $\tilde{\Psi}$ and $\tilde{\Omega}$ are set to be:

$$\underset{m\times m}{\tilde{\Psi}} = diag(\sigma_1^2,...,\sigma_m^2),$$

and

$$\underset{k\times k}{\tilde{\Omega}} = \lambda^2 \cdot diag\left(\frac{1}{\sigma_1^2},...,\frac{1}{\sigma_m^2};\frac{1}{2^2\cdot\sigma_1^2},...,\frac{1}{2^2\cdot\sigma_m^2};...;\frac{1}{p^2\cdot\sigma_1^2},...,\frac{1}{p^2\cdot\sigma_m^2};\frac{1}{\lambda^2\cdot\varepsilon}\right), \qquad (5)$$

where ε is a very small number. These parameters make the prior variance of the coefficients, $\text{var}(\mathbf{b}) = \tilde{\Psi}\otimes\tilde{\Omega}$, follow the Minnesota prior, with the one exception of the hyperparameter π, which must be equal to 1 (See Kadiyala and Karlson, 1997, or Robertson and Tallman, 1999, for more details). In practice, each parameter σ_i^2 is set to be the variance of the OLS residual from a univariate autoregressive model of order p of the variable $y_{i,t}$.

The posterior distribution of this model is also Normal-(Inverted)-Wishart, given by:

$$\underset{km\times 1}{\mathbf{b}} \mid \underset{m\times m}{\Psi}, \underset{T\times m}{Y} \sim N(\underset{km\times 1}{\bar{\mathbf{b}}}, \underset{m\times m}{\Psi} \otimes \underset{k\times k}{\bar{\Omega}}) \text{ and } \underset{m\times m}{\Psi} \mid \underset{T\times m}{Y} \sim iW(\underset{m\times m}{\bar{\Psi}}, T+\alpha), \tag{6}$$

where $\bar{\Omega} = (\tilde{\Omega}^{-1} + X'X)^{-1}$, $\bar{\mathbf{b}}$ is obtained from stacking columns of the matrix $\bar{B}$, given by:

$$\bar{B} = (\tilde{\Omega}^{-1} + X'X)^{-1}(\tilde{\Omega}^{-1}\tilde{B} + X'Y), \tag{7}$$

and $\bar{\Psi}$ is given by:

$$\bar{\Psi} = Y'Y - \bar{B}'(\tilde{\Omega}^{-1} + X'X)\bar{B} + \tilde{B}'\tilde{\Omega}^{-1}\tilde{B} + \tilde{\Psi}. \tag{8}$$

Normally, the posterior mean $\bar{\mathbf{b}}$ is used as the point estimate of the model.

With the OLS estimator in (3), the estimator of the posterior mean from (7) can be rewritten as:

$$\bar{B} = (\tilde{\Omega}^{-1} + X'X)^{-1}(\tilde{\Omega}^{-1}\tilde{B} + (X'X)\hat{B}). \tag{9}$$

The estimator in (9) looks similar to a weighted average between the prior mean $\tilde{B}$ and the OLS estimator $\hat{B}$ of the model. It is actually a shrinkage estimator that shrinks the OLS estimate toward the prior mean, which is the random walk in this case. Since λ determines the magnitude of the matrix $\tilde{\Omega}$, setting different values of λ is equivalent to assigning different relative weights to the prior information. In one extreme, if $\lambda = 0$, we give the whole weight toward the prior information. If $\lambda = +\infty$, we give the whole weight toward the OLS estimator.

Mathematically, the main problem with the OLS estimator (3) is the singularity of the matrix $X'X$. The posterior mean of the Bayesian VARs as in (9) avoids this problem by summing the diagonal matrix $\tilde{\Omega}^{-1}$ into the matrix $X'X$. This technique produces a feasible and more reliable (less variance) estimator when the number of parameters is too large relative to the number of observations.

3. Performances with Different Hyperparameter Values

The main method we use in evaluating the performance of each VAR specification is the out-of-sample assessment. We follow BGR's practice closely. The data set is of Stock and Watson (2005), which have 132 monthly macroeconomic indicators running from *January 1959* to *December 2003*.

Let $\hat{Y}^{(\mu,\lambda)}_{t+h|t} = (\hat{y}^{(\mu,\lambda)}_{1,t+h|t} \ldots \hat{y}^{(\mu,\lambda)}_{m,t+h|t})'$ denote the point estimate of the *h*-steps ahead forecast obtained from the model μ with the hyperparameter value λ. The point estimate of the one-step ahead forecast is computed from:

$$\hat{Y}^{(\mu,\lambda)}_{t+1|t}{}' = X_{t+1}{}'\overline{B}^{(\mu,\lambda)}, \tag{10}$$

where $\overline{B}^{(\mu,\lambda)}$ is the posterior mean of the coefficients from the model μ with the hyperparameter value λ. For the case of $p > h > 1$ that we consider, we can recursively construct a matrix of independent variables $X^{(\mu,\lambda)}_{t+h|t}$, given by:

$$X^{(\mu,\lambda)}_{t+h|t} = (\hat{Y}^{(\mu,\lambda)}_{t+h-1|t}{}',\ldots,\hat{Y}^{(\mu,\lambda)}_{t+1|t}{}',Y_t{}',\ldots,Y_{t+h-p}{}',1)', \tag{11}$$

using the forecasts $\hat{Y}^{(\mu,\lambda)}_{t+h-1|t},\ldots,\hat{Y}^{(\mu,\lambda)}_{t+1|t}$ and the sample observations $Y_t,\ldots,Y_{t+h-p}$. The point estimate of the *h*-steps forecast, then, is computed from:

$$\hat{Y}^{(\mu,\lambda)}_{t+h|t}{}' = X^{(\mu,\lambda)}_{t+h|t}{}'\overline{B}^{(\mu,\lambda)}. \tag{12}$$

The random walk is used as our benchmark model. The estimator can be obtained by setting λ equal to 0, which makes the *h*-steps ahead forecast from this model to be the same across all model specifications μ. We use $\hat{Y}^{(0)}_{t+h|t}$ to denote the *h*-steps ahead forecast from this benchmark model. Most of the parameter δ_i are set to be 1, except for some stationary variables specified by BGR, of which δ_i are set to be 0 (See the last column of the Appendix C).

The out-of-sample assessment is conducted for forecast horizons *h* equal to 1, 3, 6, and 12. Let t_0 and t_1 denote the position of *January 1971* and *December 2003* in the data set. For each forecast horizon *h*, we compute $\hat{Y}^{(\mu,\lambda)}_{t+h|t}$ in each period t = $t_0 - h$, …, $t_1 - h$ (396 times). The order of the VAR is *p* = 13. The parameters and posterior mean in each model for each *t* are computed from the most recent 10 years of sample observations up to time *t* (Rolling scheme, 120 observations). We set the small number ε, the parameter governing prior variances of the constant terms in the matrix $\tilde{\Omega}$ in (5), to be 10^{-10}.

The forecasting performance is measured in terms of out-of-sample Mean Squared Forecast Error (MSE). For the model μ, the value λ, the forecast horizon *h*, and the variable *i*, we have:

$$MSFE^{(\mu,\lambda)}_{i,h} = \frac{1}{t_1 - t_0 + 1} \sum_{t=t_0-h}^{t_1-h} \left(y_{i,t+h} - \hat{y}^{(\mu,\lambda)}_{i,t+h|t} \right)^2 . \tag{13}$$

The results are reported for MSFE relative to the benchmark model "Random walk with drift", given by:

$$RMSFE^{(\mu,\lambda)}_{i,h} = \frac{MSFE^{(\mu,\lambda)}_{i,h}}{MSFE^{(0)}_{i,h}} . \tag{14}$$

A number smaller than one for $RMSFE^{(\mu,\lambda)}_{i,h}$ implies that the model μ with value λ performs better than the random walk.

The variable of interest *i* are 1) employment (EMPL), measured by the number of employees on non-farm payrolls, 2) consumer price index (CPI) representing the price level, and 3) the Federal Fund Rate (FFR) representing the monetary instrument.

Following BGR, there are 4 VAR specifications μ, which are:

1. SMALL. There are only 3 variables of interest: 1) EMPL, 2) FFR, and 3) CPI.

2. CEE. This is the model of Christiano, Eichenbaum, and Evans (1999). There are 7 variables, 3 as in *SMALL*, and additionally 4) index of sensitive material prices, 5) non-borrowed reserves, 6) total reserves, and 7) M2 money stock. (I believe this should be its own paragraph, renumber 2 &3 below)

3. MEDIUM. There are 20 variables, 7 as in *CEE*, and additionally 8) personal income, 9) real consumption, 10) industrial production, 11) capacity utilization, 12) the unemployment rate, 13) housing starts, 14) producer price index, 15) personal consumption expenditures price deflator, 16) average hourly earnings, 17) M1 money stock, 18) Standard and Poor's price index, 19) Yields on 10-year US Treasury bond, and 20) effective exchange rate.

4. LARGE. This specification includes all indicators in the data set, except for the spot market price index of all commodities (PSCCOM).

We report our first out-of-sample assessment result in Table 1, using the same hyperparameter values as in BGR. That is $\lambda = \infty$ for μ= *SMALL*, $\lambda = 0.262$ for μ= *CEE*, $\lambda = 0.108$ for μ= *MEDIUM*, and $\lambda = 0.035$ for μ= *LARGE*. This result is qualitatively similar to Table 1 of BGR. It can be seen clearly that larger models perform better than smaller ones.

Table 1: BVARs different λ, Out-of-Sample Relative MSFE, 1971 – 2003

		SMALL	*CEE*	*MEDIUM*	*LARGE*
$h = 1$	EMPL	1.02	0.65	0.54	0.45
	FFR	1.65	0.90	0.79	0.75
	CPI	0.81	0.55	0.51	0.51
$h = 3$	EMPL	0.85	0.63	0.50	0.37
	FFR	1.57	1.12	0.96	0.92
	CPI	0.60	0.43	0.40	0.40
$h = 6$	EMPL	0.90	0.79	0.66	0.51
	FFR	1.84	1.30	1.31	1.24
	CPI	0.59	0.44	0.37	0.40
$h = 12$	EMPL	0.84	0.96	0.87	0.81
	FFR	2.48	1.49	1.56	1.80
	CPI	0.74	0.60	0.44	0.45
λ		∞	0.262	0.108	0.035

BGR assign hyperparameter values to keep the in-sample fit of all models in the pre-evaluation period to be the same, for the forecast horizon *h* = 1. Specifically, let T_0 denote the position of *December 1969* in the data set. Define the in-sample 1-step ahead mean squared forecast errors (*msfe*) for a model μ, a hyperparameter value λ, and a variable *i* as:

$$msfe_i^{(\mu,\lambda)} = \frac{1}{T_0 - p - 1} \sum_{t=p}^{T_0 - 1} \left(\hat{y}_{i,t+1|t}^{(\mu,\lambda)} - y_{i,t+1} \right)^2 .$$

Note that $\hat{y}_{i,t+1|t}^{(\mu,\lambda)}$ is the in-sample forecast (Estimated value) for $y_{i,t+1}$ within the period from *January 1960* (*t* = 1) to *December 1969* (*t* = T_0).

Next, estimate the unrestricted OLS VAR of the *SMALL* model using the data from *January 1960* to *December 1969*, and figure out the in-sample fit (*Fit*), given by:

$$Fit = \frac{1}{3} \sum_{i \in I} \frac{msfe_i^{(\mu,\lambda)}}{msfe_i^{(0)}} \Bigg|_{\mu = SMALL, \lambda = +\infty} ,$$

where $I = \{\text{EMPL}, \text{FFR}, \text{CPI}\}$ is the set of variables of interest.

At last, for each model $\mu \neq SMALL$, determine using a grid search the hyperparameter $\lambda^{(\mu,Fit)}$ that gives the in-sample fit of the model closest to the in-sample fit of the unrestricted OLS VAR. Specifically, the hyperparameter $\lambda^{(\mu,Fit)}$ can be defined as:

$$\lambda^{(\mu,Fit)} = \arg\min_{\lambda} \left| Fit - \frac{1}{3} \sum_{i \in I} \frac{msfe_i^{(\mu,\lambda)}}{msfe_i^{(0)}} \right| .$$

We see the way BGR set the hyperparameter value biases against small models. First, note that the *SMALL* Bayesian VAR of BGR is actually the unrestricted OLS VAR. This is because the hyperparameter of the model is set at $\lambda = +\infty$. The *SMALL* model does not benefit from shrinkage estimation at all. Next, observe that larger models will be assigned lower values of the hyperparameter λ. This is a usual result as a larger OLS model provides a better in-sample fit to the sample observations. To set the in-sample fit at a given level, this model must be pulled away more from its OLS estimate. However, since the shrinkage estimator improves the forecasting performance of a model by avoiding the problem of overfitting into the sample observations[1], this way of assigning the hyperparameter values provides more benefits to larger models.

To show this empirically, we set up a new out-of-sample assessment that assigns the same hyperparameter value across all model specifications. Each hyperparameter value $\lambda = 0.035$, 0.108, and 0.262 is applied to all specifications in this assessment. Everything else stays the same. Table 2 reports the relative MSFE under this new assessment.

Comparing Table 2 to Table 1, we can see an obvious improvement in the performance of small models. Even the smallest models can benefit from shrinkage estimation. The *SMALL* is a 3-variable VAR with 13 lags, which results in 40 coefficients to be estimated per equation including the constant term. It is estimated with 120 observations each time. The smallest amount

[1] Zha (1998) provides a good discussion on this point.

of shrinkage in this experiment, $\lambda = 0.262$, can remarkably improve the forecasting performances of this model.

The obvious improvement of the forecasting performances for larger models disappears. This point is apparent for the hyperparameter values $\lambda = 0.108$ and $\lambda = 0.262$. Especially noteworthy is that for the case of fixing the value at $\lambda = 0.262$, the *SMALL* model outperforms other larger models in the overall picture. This suggests that the value of hyperparameter λ must be chosen more carefully.

4. Performances with Suitable Hyperparameter Values

In this section, we consider a procedure for choosing a "suitable" hyperparameter value for each model based on a training sample. We choose the value that leads to the minimum relative MSFE of the variables of interest in an out-of-sample assessment. The observations up to *December 1980* are employed to figure out each suitable hyperparameter value for a given VAR specification μ and a given forecasting span *h*. After that we will assess each model with this hyperparameter value, using our evaluation period from *January 1981* to *December 2003*.

Table 2: BVARs same λ, Out-of-Sample Relative MSFE, 1971 – 2003

λ = 0.035		*SMALL*	*CEE*	*MEDIUM*	*LARGE*
h = 1	EMPL	0.64	0.61	0.52	0.45
	FFR	1.00	0.95	0.87	0.75
	CPI	0.57	0.51	0.51	0.51
h = 3	EMPL	0.63	0.56	0.48	0.37
	FFR	1.06	1.03	1.02	0.92
	CPI	0.47	0.38	0.38	0.40
h = 6	EMPL	0.73	0.63	0.58	0.51
	FFR	1.11	1.11	1.28	1.24
	CPI	0.46	0.34	0.34	0.40
h = 12	EMPL	0.93	0.71	0.72	0.81
	FFR	1.22	1.27	1.56	1.80
	CPI	0.51	0.40	0.38	0.45

λ = 0.108		*SMALL*	*CEE*	*MEDIUM*	*LARGE*
h = 1	EMPL	0.54	0.59	0.54	0.51
	FFR	0.96	0.86	0.79	0.75
	CPI	0.55	0.51	0.51	0.55
h = 3	EMPL	0.49	0.55	0.50	0.40
	FFR	1.08	0.94	0.96	0.94
	CPI	0.47	0.39	0.40	0.46
h = 6	EMPL	0.55	0.66	0.66	0.54
	FFR	1.18	1.03	1.31	1.33
	CPI	0.48	0.37	0.37	0.47
h = 12	EMPL	0.65	0.76	0.87	0.96
	FFR	1.29	1.21	1.56	1.86
	CPI	0.55	0.47	0.44	0.59

λ = 0.262		*SMALL*	*CEE*	*MEDIUM*	*LARGE*
h = 1	EMPL	0.57	0.65	0.66	0.65
	FFR	0.92	0.90	0.85	0.82
	CPI	0.55	0.55	0.54	0.62
h = 3	EMPL	0.53	0.63	0.63	0.47
	FFR	1.14	1.12	1.04	1.05
	CPI	0.48	0.43	0.44	0.51
h = 6	EMPL	0.60	0.79	0.82	0.62
	FFR	1.29	1.30	1.51	1.49
	CPI	0.50	0.44	0.39	0.52
h = 12	EMPL	0.65	0.96	1.10	1.09
	FFR	1.50	1.49	1.85	1.96
	CPI	0.59	0.60	0.49	0.69

In searching for a suitable hyperparameter value, let τ_0 and τ_1 denote the position of *January 1971* and *December 1980* in the data set, respectively. For a forecasting span $h \in \{1,3,6,12\}$ and an arbitrary hyperparameter value $\tilde{\lambda}$, we can compute a forecast $\hat{Y}^{(\mu,\tilde{\lambda})}_{\tau+h|\tau}$ for each period $\tau = \tau_0 - h, ..., \tau_1 - h$ (120 times), using the same setting as in the previous section (*p* = 13, rolling scheme with 120 observations, $\varepsilon = 10^{-10}$). These point forecasts can be used to compute $MSFE^{(\mu,\tilde{\lambda})}_{i,h}$ and $RMSFE^{(\mu,\tilde{\lambda})}_{i,h}$ from (13) and (14), for each variable of interest *i*. Let $TV^{(\mu,\tilde{\lambda})}_h \equiv \sum_{i \in I} RMSFE^{(\mu,\tilde{\lambda})}_{i,h}$ denote our target variable, which is the sum of relative MSFE of our three variables of interest. Note that each relative MSFE does not depend on the unit of measurement of each variable, since it is a (xx ? xx) relative term. We find the suitable hyperparameter value λ^μ_h for each specification μ and each forecast horizon *h* using a grid search such that:

$$\lambda^\mu_h = \arg\min_{\tilde{\lambda}} TV^{(\mu,\tilde{\lambda})}_h . \tag{15}$$

Since there is no natural upper bound for the hyperparameter λ, a good grid search for λ^μ_h should cover a wide range of possible values between 0 and $+\infty$. To avoid this, we calculate the derivative $\partial TV^{(\mu,\tilde{\lambda})}_h \big/ \partial\tilde{\lambda}$ to help search for each hyperparameter value λ^μ_h. This allows us to search for the value λ^μ_h in steps, and helps reduce the task.

From (13) and (14), we have:

$$\frac{\partial TV^{(\mu,\tilde{\lambda})}_h}{\partial\tilde{\lambda}} = \frac{\partial MSFE^{(\mu,\tilde{\lambda})}_{\mathrm{EMPL},h} \big/ \partial\tilde{\lambda}}{MSFE^{(0)}_{\mathrm{EMPL},h}} + \frac{\partial MSFE^{(\mu,\tilde{\lambda})}_{\mathrm{CPI},h} \big/ \partial\tilde{\lambda}}{MSFE^{(0)}_{\mathrm{CPI},h}} + \frac{\partial MSFE^{(\mu,\tilde{\lambda})}_{\mathrm{FFR},h} \big/ \partial\tilde{\lambda}}{MSFE^{(0)}_{\mathrm{FFR},h}}, \tag{16}$$

where for each variable $i \in I$, $\left\{ \text{EMPL< CPI, FFR} \right\}$,

$$\begin{aligned}\frac{\partial MSFE_{i,h}^{(\mu,\tilde{\lambda})}}{\partial \tilde{\lambda}} &= \frac{1}{\tau_1 - \tau_0 + 1} \sum_{t=\tau_0 - h}^{\tau_1 - h} \frac{\partial}{\partial \tilde{\lambda}} \left(y_{i,t+h} - \hat{y}_{i,t+h|t}^{(\mu,\tilde{\lambda})} \right)^2, \\ &= \frac{1}{\tau_1 - \tau_0 + 1} \sum_{t=\tau_0 - h}^{\tau_1 - h} -2 \left(y_{i,t+h} - \hat{y}_{i,t+h|t}^{(\mu,\tilde{\lambda})} \right) \frac{\partial \hat{y}_{i,t+h|t}^{(\mu,\tilde{\lambda})}}{\partial \tilde{\lambda}}.\end{aligned} \tag{17}$$

Given an $m \times n$ matrix *Z*, we use $\partial Z / \partial \tilde{\lambda}$ to denote the gradient matrix of *Z* with respect to $\tilde{\lambda}$. The gradient matrix $\partial Z / \partial \tilde{\lambda}$ has the same dimension as *Z* with $\partial z_{ij} / \partial \tilde{\lambda}$ as its *ij*-th element, where z_{ij} is the *ij*-th element of *Z*. This is the same for a gradient vector $\partial \mathbf{z} / \partial \tilde{\lambda}$ of an $m \times 1$ vector **z**. The value of $\partial \hat{y}_{i,t+h|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda}$ in (17) can be taken from the gradient vector $\partial \hat{Y}_{t+h|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda}$, which, according to (12), can be written as:

$$\left(\partial \hat{Y}_{t+h|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} \right)' = \left(\partial X_{t+h|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} \right)' \bar{B}^{(\mu,\tilde{\lambda})} + X_{t+h|t}^{(\mu,\tilde{\lambda})}{}' \left(\partial \bar{B}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} \right), \tag{18}$$

where $\partial \bar{B}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda}$ is given by[2]:

$$\partial \bar{B}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} = \left(\tilde{\Omega}^{(\mu,\tilde{\lambda})\,-1} + X^{(\mu)}{}' X^{(\mu)} \right)^{-1} \left(\partial \tilde{\Omega}^{(\mu,\tilde{\lambda})\,-1} / \partial \tilde{\lambda} \right) \left(\tilde{B}^{(\mu)} - \bar{B}^{(\mu,\tilde{\lambda})} \right). \tag{19}$$

We can compute the gradient vector $\partial \hat{Y}_{t+h|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda}$ recursively, using (18) and, according to (10) and (11), the following equations:

$$\left(\partial \hat{Y}_{t+1|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} \right)' = X_{t+1}{}' \left(\partial \bar{B}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} \right), \tag{20}$$

and

$$\partial X_{t+h|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} = \left(\left(\partial \hat{Y}_{t+h-1|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} \right)' \ \ldots \ \left(\partial \hat{Y}_{t+1|t}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda} \right)' \ \underset{1 \times (k - mh + m)}{\mathbf{0}} \right)'. \tag{21}$$

[2] See the Appendix A for the derivation of $\partial \bar{B}^{(\mu,\tilde{\lambda})} / \partial \tilde{\lambda}$ in (19).

We perform a grid search to find the values of λ_h^{μ} for each forecasting horizon *h* and each model specification μ. We search for λ_h^{μ} with 3 decimal places, which makes our grid search perform 4 steps as follow[3]:

1. Calculate the values of target variables $TV_h^{(\mu,\tilde{\lambda})}$ and gradient $\partial TV_h^{(\mu,\tilde{\lambda})} \big/ \partial\tilde{\lambda}$ for each of 11 values of $\tilde{\lambda}$, which are 0.001 and 1,2,…,10. Figure out the possible region of the hyperparameter value λ_h^{μ} [4]. Let λ_{s1} denote the lower bound of this region.

2. Calculate the values of target variables $TV_h^{(\mu,\tilde{\lambda})}$ and gradient $\partial TV_h^{(\mu,\tilde{\lambda})} \big/ \partial\tilde{\lambda}$ for each of 9 values of $\tilde{\lambda}$, which are λ_{s1}+0.1, λ_{s1}+0.2,…, λ_{s1}+0.9. Figure out the possible region of the hyperparameter value λ_h^{μ}. Let λ_{s2} denote the lower bound of this region.

3. Calculate the values of target variables $TV_h^{(\mu,\tilde{\lambda})}$ and gradient $\partial TV_h^{(\mu,\tilde{\lambda})} \big/ \partial\tilde{\lambda}$ for each of 9 values of $\tilde{\lambda}$, which are λ_{s2}+0.01, λ_{s2}+0.02,…, λ_{s2}+0.09. Figure out the possible region of the hyperparameter value λ_h^{μ}. Let λ_{s3} denote the lower bound of this region.

4. Calculate the values of target variables $TV_h^{(\mu,\tilde{\lambda})}$ for 9 values of $\tilde{\lambda}$, which are λ_{s3}+0.001, λ_{s3}+0.002,…, λ_{s3}+0.009. The suitable hyperparameter value λ_h^{μ} is the one associated with the minimum value of $TV_h^{(\mu,\tilde{\lambda})}$ from these 4 steps.

Table 3 reports the optimal hyperparameter λ_h^{μ} with the associated values of target $TV_h^{(\mu,\tilde{\lambda})}$ and gradient, for each forecast horizon *h* and each model specification μ. The details of grid search can be found in Appendix B.

[3] It is possible that our grid search may not return the optimal hyperparameter λ_h^{μ} as defined in (15), if the function $TV_h^{(\mu,\tilde{\lambda})}$ is not smooth. Regarding this problem, we have tried minimizing the function with respect to the value of $\tilde{\lambda}$ for each forecast horizon h = 1,3,6,12 of the SMALL model, using the add-on application OPTMUM in GAUSS. It returns a similar result to our grid search. The use of this program, however, is not practical for larger models, as it consumes a lot of processing time even for our smallest model.

[4] This should be the region that has a negative gradient at its lower bound. This tells that the values in the region will generate smaller values of $TV_h^{(\mu,\tilde{\lambda})}$ than one at the lower bound.

Table 3: λ_h^μ from grid search and $TV_h^{(\mu,\tilde{\lambda})}$, 1971 – 1980

		SMALL	*CEE*	*MEDIUM*	*LARGE*
$h = 1$	λ_h^μ	0.130	0.129	0.096	0.053
	$TV_h^{(\mu,\tilde{\lambda})}$	1.950	1.849	1.665	1.598
	gradient	0.008	0.001	–0.000	0.029
$h = 3$	λ_h^μ	0.111	0.143	0.117	0.072
	$TV_h^{(\mu,\tilde{\lambda})}$	1.876	1.739	1.695	1.614
	gradient	0.001	0.004	0.024	–0.008
$h = 6$	λ_h^μ	0.130	0.134	0.017	0.059
	$TV_h^{(\mu,\tilde{\lambda})}$	1.887	1.852	2.191	2.288
	gradient	0.007	0.004	–0.350	–0.033
$h = 12$	λ_h^μ	0.102	0.049	0.020	0.006
	$TV_h^{(\mu,\tilde{\lambda})}$	1.912	1.896	2.254	2.483
	gradient	–0.000	–0.020	–0.672	–1.646

Using the suitable values λ_h^μ from Table 3, we perform the out-of-sample assessment. Let t_0 and t_1 denote the position of *January 1981* and December *2003*, respectively, in the data set. We compute the forecast $\hat{Y}_{t+h|t}^{(\mu,\lambda_h^\mu)}$ in each period $t = t_0 - h, \ldots, t_1 - h$ (276 times) with VAR of order p = 13, using the most recent 10 years of observations (Rolling scheme, 120 observations), and the parameter ε at 10^{-10}. These forecasts are used to calculate the relative MSFE in (14), for 3 variables of interest i = EMPL, FFR, and CPI. Table 4 reports the result of this assessment, with the associated values of $TV_h^{(\mu,\lambda_h^\mu)} = \sum_{i\in I} RMSFE_i^{(\mu,\lambda_h^\mu)}$ and λ_h^μ.

Since our evaluation period has been changed from the previous section, we also construct Table 5 for the purpose of comparison. In this table, we use the same setting as in Table 1 of the previous section, but the evaluation period has been changed to one from *January 1981* to *December 2003*.

According to Table 4, the *LARGE* model gives the best overall performance. This supports the finding of BGR that larger Bayesian VARs perform better than smaller models in forecasting the three key macroeconomic variables. However, we see a significant difference between Table 4 and Table 5.

Table 4: BVARs with λ_h^μ, Out-of-Sample Relative MSFE, 1981 – 2003

		SMALL	*CEE*	*MEDIUM*	*LARGE*
$h = 1$	EMPL	0.53	0.62	0.53	0.49
	FFR	0.96	0.85	0.93	0.80
	CPI	0.62	0.60	0.57	0.53
	$TV_h^{(\mu,\lambda_h^\mu)}$	2.104	2.067	2.025	1.822
	λ_h^μ	0.130	0.129	0.096	0.053
$h = 3$	EMPL	0.42	0.56	0.44	0.37
	FFR	1.23	1.06	1.13	0.95
	CPI	0.59	0.52	0.54	0.51
	$TV_h^{(\mu,\lambda_h^\mu)}$	2.234	2.145	2.107	1.830
	λ_h^μ	0.111	0.143	0.117	0.072
$h = 6$	EMPL	0.53	0.77	0.63	0.49
	FFR	1.47	1.12	1.17	1.05
	CPI	0.62	0.51	0.43	0.50
	$TV_h^{(\mu,\lambda_h^\mu)}$	2.612	2.391	2.225	2.045
	λ_h^μ	0.130	0.134	0.017	0.059
$h = 12$	EMPL	0.72	0.91	0.82	0.69
	FFR	1.47	1.24	1.75	1.75
	CPI	0.78	0.54	0.47	0.52
	$TV_h^{(\mu,\lambda_h^\mu)}$	2.966	2.696	3.045	2.970
	λ_h^μ	0.102	0.049	0.020	0.006

Table 5: BVARs different λ, Out-of-Sample Relative MSFE, 1981 – 2003

		SMALL	*CEE*	*MEDIUM*	*LARGE*
$h = 1$	EMPL	0.81	0.68	0.54	0.47
	FFR	1.71	0.99	0.94	0.78
	CPI	0.83	0.64	0.57	0.54
	$TV_h^{(\mu,\lambda)}$	3.360	2.313	2.049	1.794
$h = 3$	EMPL	0.67	0.65	0.43	0.32
	FFR	1.74	1.43	1.12	0.89
	CPI	0.73	0.58	0.53	0.48
	$TV_h^{(\mu,\lambda)}$	3.132	2.649	2.077	1.689
$h = 6$	EMPL	0.89	0.95	0.62	0.45
	FFR	2.44	1.58	1.36	1.00
	CPI	0.77	0.59	0.48	0.47
	$TV_h^{(\mu,\lambda)}$	4.105	3.117	2.461	1.925
$h = 12$	EMPL	1.04	1.33	0.93	0.82
	FFR	3.18	1.63	1.40	1.64
	CPI	1.04	0.80	0.51	0.56
	$TV_h^{(\mu,\lambda)}$	5.266	3.761	2.840	3.014
λ		∞	0.262	0.108	0.035

The results of Table 5 seem to indicate that adding more variables into the VAR helps to significantly improve its forecasting performances. The models with 7 and 20 variables perform much better than the 3-variables model. Since Bayesian or shrinkage estimation allows us to use all available information in making forecasts, adding as many data as possible like the *LARGE* model helps further improve the forecasting performances.

However, Table 4 shows that this impression is false. This is a result of allowing no shrinkage at all for the *SMALL* model. If we use Bayesian or shrinkage estimation with the *SMALL* model, the improvement of larger VARs over the 3-variables VAR becomes minimal. Specifically, the 7-variables and 20-variables models do not seem to have a clear edge over the 3-variables model, and the improvement of the 131-variables VAR is much less pronounced than what Table 5 implies. This is also the case after we have tried to make the most out of each model given our pre-evaluation period.

5. Repeated Calculations of Hyperparameter Values

It can be the case that the optimal hyperparameter value λ_h^μ varies with time. Allowing some changes for the value may improve the forecasting performance of each Bayesian VAR. In this section, we allow this change every 10 years. We repeat our practice in the previous section of finding the suitable hyperparameter value after we have an additional 10 years of observations.

Table 6 reports the suitable hyperparameter value λ_h^μ with the associated values of the target variable and gradient for each forecast horizon *h* and each model μ, using the observations from *January 1971* to *December 1990*. Table 7 reports the same values, using the observations from *January 1971* to *December 2000*. The suitable hyperparameter values reported in Table 6 and Table 7 look different from the ones in Table 3 of the previous section. However, the values are relatively similar in these two tables.

Table 6: λ_h^μ from grid search and $TV_h^{(\mu,\tilde{\lambda})}$, 1971 – 1990

		SMALL	*CEE*	*MEDIUM*	*LARGE*
$h = 1$	λ_h^μ	0.164	0.102	0.078	0.044
	$TV_h^{(\mu,\tilde{\lambda})}$	2.003	1.920	1.793	1.666
	gradient	–0.002	0.007	0.010	–0.069
$h = 3$	λ_h^μ	0.089	0.085	0.066	0.048
	$TV_h^{(\mu,\tilde{\lambda})}$	2.044	1.857	1.849	1.691
	gradient	–0.005	0.003	–0.036	–0.063
$h = 6$	λ_h^μ	0.059	0.066	0.022	0.043
	$TV_h^{(\mu,\tilde{\lambda})}$	2.195	1.986	2.246	2.199
	gradient	–0.030	0.014	0.033	–0.002
$h = 12$	λ_h^μ	0.073	0.055	0.047	0.005
	$TV_h^{(\mu,\tilde{\lambda})}$	2.407	2.227	2.691	2.772
	gradient	0.018	0.064	0.126	12.717

Table 7: λ_h^μ from grid search and $TV_h^{(\mu,\tilde{\lambda})}$, 1971 – 2000

		SMALL	*CEE*	*MEDIUM*	*LARGE*
$h = 1$	λ_h^μ	0.168	0.101	0.077	0.043
	$TV_h^{(\mu,\tilde{\lambda})}$	2.030	1.963	1.828	1.704
	gradient	–0.001	–0.007	0.015	–0.008
$h = 3$	λ_h^μ	0.098	0.083	0.062	0.046
	$TV_h^{(\mu,\tilde{\lambda})}$	2.045	1.877	1.852	1.699
	gradient	–0.002	–0.018	–0.018	0.096
$h = 6$	λ_h^μ	0.071	0.064	0.022	0.041
	$TV_h^{(\mu,\tilde{\lambda})}$	2.199	2.004	2.212	2.176
	gradient	0.000	–0.010	–0.225	–0.028
$h = 12$	λ_h^μ	0.086	0.054	0.043	0.005
	$TV_h^{(\mu,\tilde{\lambda})}$	2.434	2.276	2.690	2.724
	gradient	–0.017	–0.004	0.134	4.146

Next, we use the hyperparameter values from Table 3, Table 6, and Table 7 in assessing the out-of-sample forecasting performances of our Bayesian VARs. The values from Table 3 are used to make forecasts from *January 1981* to *December 1990*. Ones from Table 6 are used for the forecasts from *January 1991* to *December 2000*, and ones from Table 7 for *January 2001* to *December 2003*. Table 8 reports the values of out-of-sample relative MSFE from this exercise. The variable TV_h^μ represents the sum of relative MSFE of our variables of interest for model μ and forecast horizon *h*. Comparing the results in Table 8 and Table 4, we can see that our exercise only marginally improves the forecasting performances of the models.

Table 8: BVARs with Varied λ, Out-of-Sample Relative MSFE, 1981 – 2003

		SMALL	*CEE*	*MEDIUM*	*LARGE*
$h = 1$	EMPL	0.53	0.62	0.53	0.49
	FFR	0.96	0.85	0.93	0.80
	CPI	0.62	0.60	0.57	0.53
	TV_h^μ	2.102	2.063	2.025	1.821
$h = 3$	EMPL	0.43	0.57	0.45	0.37
	FFR	1.23	1.05	1.12	0.95
	CPI	0.59	0.51	0.52	0.50
	TV_h^μ	2.241	2.133	2.090	1.819
$h = 6$	EMPL	0.58	0.79	0.61	0.48
	FFR	1.46	1.08	1.16	1.05
	CPI	0.60	0.49	0.44	0.50
	TV_h^μ	2.650	2.362	2.202	2.036
$h = 12$	EMPL	0.75	0.91	0.78	0.71
	FFR	1.46	1.24	1.73	1.76
	CPI	0.77	0.54	0.51	0.52
	TV_h^μ	2.987	2.693	3.018	2.987

6. An Updating Scheme for the Hyperparameter

Another way to allow changes in hyperparameter values is to use an updating scheme that is sensitive to previous forecasting performances of the model. In this section, we apply an updating scheme that makes use of each additional observation in determining whether to change the hyperparameter value of a model. Such adaptive schemes will only improve forecasting performance if the underlying data generating process (DGP) is changing through time. What the "optimal" adaptive scheme will be depends on how the underlying DGP is changing over time.

Although there is a strong belief that there has been structural change in the economic system during our sample period, there is no precise information about how the parameters have changed. Therefore, instead of making an arbitrary assumption about what mechanisms may be governing such changes and then driving the optimal adaptive scheme for that mechanism, we consider an adaptive scheme that makes good sense to us. From the practice of forecasting, we know that adaptive schemes that give very high weight to new information often chase noises and do not perform very well. Hence, we consider the following scheme.

Let t_0 and t_1 represent the positions of *January 1981* and *December 2003*, respectively. We start using the hyperparameter value of each model and each forecast horizon from Table 3. We use λ^{μ}_{h,t_0} to denote this initial hyperparameter value. Let $\lambda^{\mu}_{h,T}$ denote the value used in a given period *T*. At the start of each period $T \in [t_0, t_1]$, we use a model μ in making a forecast $\hat{Y}^{(\mu,\lambda^{\mu}_{h,T})}_{T|T-h}$. At the end of the period *T*, after realizing the actual data Y_T, we calculate the square forecast error from the Bayesian VAR μ:

$$SFE^{(\mu,\lambda^{\mu}_{h,T})}_{i,h;T} \equiv (y_{i,T} - \hat{y}^{\mu,\lambda^{\mu}_{h,T}}_{T|T-h})^2, \tag{22}$$

as well as the square forecast error from the benchmark model $SFE^{(0)}_{i,h;T}$ for each variable of interest $i \in I$.

We also calculate at this point in time the indicators:

$$INDC_{h;T}^{(\mu,\lambda)} \equiv \sum_{i\in I} \frac{SFE_{i,h;T}^{(\mu,\lambda)}}{\sum_{t=t_0}^{T} SFE_{i,h;t}^{(0)} + (276+t_0-T)SFE_{i,h;T}^{(0)}}, \tag{23}$$

for 3 values of the hyperparameter λ, which are $\lambda_{h,T}^{\mu}$, $\lambda_{h,T}^{\mu}$+0.001, and $\lambda_{h,T}^{\mu}$-0.001 [5]. We use the indicator $INDC_{h;T}^{(\mu,\lambda)}$ to approximate the marginal increase in the sum of relative MSFE from using different values of λ at time *T*. Observe that the term $\sum_{t=t_0}^{T} SFE_{i,h;t}^{(0)}$ in the denominator increases as *T* increases. We put the term $(276+t_0-T)SFE_{i,h;T}^{(0)}$ [6] into the denominator as well to make the value of $INDC_{h;T}^{(\mu,\lambda)}$ relatively stable along the time *T*. Otherwise, the value $\lambda_{h,T}^{\mu}$ will experience greater fluctuations for a small *T* and be very stable for larger *T*, if we fix a constant threshold as in the following .

Among these 3 hyperparameter values, we first choose the one that gives the minimum value of $INDC_{h;T}^{(\mu,\lambda)}$. If it is $\lambda = \lambda_{h,T}^{\mu}$, we also use this value as $\lambda_{h,T+1}^{\mu}$ in the next period. Otherwise, for $\lambda \in \{\lambda_{h,T}^{\mu}+0.001, \lambda_{h,T}^{\mu}-0.001\}$, if ($INDC_{h;T}^{(\mu,\lambda_{h,T}^{\mu})}$ – $INDC_{h;T}^{(\mu,\lambda)}$) is higher than 0.0001 we use this new value as $\lambda_{h,T+1}^{\mu}$ in the next period. Observe that we can increase the fluctuation of the hyperparameter value $\lambda_{h,T}^{\mu}$ by increasing the step size (Currently at 0.001) and lowering the threshold value (Currently at 0.0001). Actually, we have made some experiments with a range of threshold values and step sizes. The settings reported here yields the best results. Note also that in this process we use the information up to period *T* to figure out the hyperparameter value $\lambda_{h,T+1}^{\mu}$ that will be applied in the next time period *T*+1.

[5] We use step size equal to 0.001 in every case, except for the case of *LARGE* model with *h* = 12 that we use step size at 0.0005.

[6] Recall that 276 = $t_1 - t_0 - 1$ is the total number of repetitions in our out-of-sample exercise.

At the end of the exercise, we calculate relative MSFE for each variable of interest $i \in I$ from the square forecast errors $SFE_{i,h;T}^{(\mu,\lambda_{h,T}^{\mu})}$ calculated at the start of each period T. The relative MSFE can be written as:

$$RMSFE_{i,h}^{(\mu)} = \frac{\sum_{t=t_0}^{t_1} SFE_{i,h;t}^{(\mu,\lambda_{h,t}^{\mu})}}{\sum_{t=t_0}^{t_1} SFE_{i,h;t}^{(0)}}. \tag{24}$$

Table 9 reports the relative MSFE from this exercise.

Comparing Table 9 to Table 4, there is just a small improvement to the forecasting performance of each Bayesian VAR from this exercise. This improvement, however, does not affect our finding in Section 4 that the forecasting performances of the larger models are not impressively better than is the case in the smallest model.

7. Conclusion

Bayesian or shrinkage estimation allows us to use all available information to forecast key economic indicators. BGR show us this point using the US data. The results of BGR, similar to our Table 1 or Table 5, implicitly imply that a 3-variables VAR is grossly inadequate.

However, this impression is false and is a result of their practice of not allowing any shrinkage in the 3-variables model. This 3-variables VAR has 13 lags, estimated using 120 observations. We have shown that if we use a shrinkage estimator for this 3-variables model with an appropriate hyperparameter value, the improvement of larger models will be minimal. Specifically, the 7-variables and 20-variables models considered in BGR do not seem to have a clear edge over the 3-variables model, and the improvement of the 131-variables model is much less pronounced than what BGR implies.

Table 9: BVARs with Varied λ, Out-of-Sample Relative MSFE, 1981 – 2003

		SMALL	*CEE*	*MEDIUM*	*LARGE*
h = 1	EMPL	0.53	0.62	0.52	0.48
	FFR	0.95	0.84	0.91	0.78
	CPI	0.61	0.60	0.57	0.53
	TV_h^{μ}	2.098	2.053	2.003	1.798
	$\lambda_{h,0}^{\mu}$	0.130	0.129	0.096	0.053
h = 3	EMPL	0.42	0.56	0.43	0.36
	FFR	1.23	1.04	1.12	0.91
	CPI	0.59	0.52	0.53	0.48
	TV_h^{μ}	2.237	2.124	2.079	1.747
	$\lambda_{h,0}^{\mu}$	0.111	0.143	0.117	0.072
h = 6	EMPL	0.53	0.77	0.44	0.47
	FFR	1.47	1.11	1.04	1.02
	CPI	0.62	0.50	0.50	0.49
	TV_h^{μ}	2.613	2.383	1.978	1.975
	$\lambda_{h,0}^{\mu}$	0.130	0.134	0.017	0.059
h = 12	EMPL	0.72	0.92	0.81	0.57
	FFR	1.47	1.24	1.63	1.54
	CPI	0.78	0.54	0.46	0.44
	TV_h^{μ}	2.966	2.698	2.908	2.548
	$\lambda_{h,0}^{\mu}$	0.102	0.049	0.020	0.006

We try allowing for time-varying hyperparameter values as well, but the result we have found so far is that the time-varying scheme only marginally improves the performance of each model. It does not change our previous conclusion either.

In this study, we also demonstrate a way to calculate the suitable hyperparameter value for each model specification with a given forecast horizon. The value is chosen based on the out-of-sample forecasting performances in the test period, which is a part of the pre-evaluation period. This process takes time for the *LARGE* model. The estimation of the LARGE model involves calculating for the inverse of matrix of dimension $(1{,}704 \times 1{,}704)$. Since we have to estimate the model 120 times for each value of λ and each forecast horizon *h* in the grid search shown in the Appendix B, it costs us about 3 days for each of the 4 steps in the search under the computation of a Pentium Core 2 processor. The whole process of the grid search, which is composed of 4 steps, requires about two weeks.

In practice of course, we need to figure this suitable hyperparameter value just once. We think that this is the process that should be taken rather than depending on an arbitrarily chosen value. Moreover, as can be observed from Table 6 and Table 7, the value tends to be stable for a long enough series as in the case of US data.

It would be convenient if we could figure out some patterns of changes in the suitable hyperparameter values of the Bayesian VARs. For example, the values may decrease for longer forecast horizons or bigger model specifications. Our results so far have not shown any obvious pattern. However, a more thorough investigation in this direction is still interesting.

Another interesting way to deal with the hyperparameter value is to figure out a good updating scheme. The scheme that can calculate a suitable hyperparameter value after every additional realization looks very attractive for actual forecasting excercies. Unfortunately, this hyperparameter has a non-linear relationship with the forecasting performances of the model. One might have to depend on a relatively complicated framework to figure out an optimal updating scheme for the model.

References

Bańbura, M., D. Giannone, and L. Reichlin (2008). Large Bayesian VARs. Working Paper Series 966, European Central Bank.

Bernanke, B.S., 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.

Bernanke, B.S., and J. Boivin (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics* 50(3), 525 – 546.

Christiano, L.J., M. Eichenbaum, and C.L. Evans (1999). MNonetary policy shocks: What have we learned and to what end? In J.B. Taylor and M. Woodford (Eds.), *Handbook of Macroeconomics*, Volume 1, Chapter 2, pp. 65 – 148. Elsevier.

D'Agostino, A. and D. Giannone (2007). Comparing alternative predictors based on large-panel factor models. CEPR Discussion Papers 6564, Center of Economic Policy Research.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics* 82(4), 540 – 554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2003). Do financial variables help forecasting inflation and real activity in the Euro area? *Journal of Monetary Economics* 50(6), 1243 – 1255.

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.

Litterman, R.B. (1986). Forecasting with Bayesian vector autoregressions – Five years of experience. *Journal of Business and Economic Statistics* 4(1), 25 – 38.

Magnus, J.R. and H. Neudecker (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (2nd ed.). John Wiley & Sons.

Robertson, J.C. and E.W. Tallman (1999). Vector autoregressions: Forecasting and reality. Federal Reserve Bank of Atlanta *Economic Review* 84(1), 4 – 18.

Stock, J.H. and M.W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of American Statistical Association* 97, 1167 – 1179.

Stock, J.H. and M.W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20(2), 147 – 162.

Stock, J.H. and M.W. Watson (2005). Implications of dynamic factor models for VAR analysis. NBER Working Papers 11467, National Bureau of Economic Research.

Zha, T. (1998). A dynamic multivariate model for use in formulating policy. Federal Reserve Bank of Atlanta *Economic Review* 83(1), 16 – 29.

Appendices

A. Gradient Matrix of the Coefficients

To simplify the notation, let $Z \equiv \tilde{\Omega}^{-1} + X'X$. The posterior mean $\overline{B}$ can be written as:

$$\overline{B} = Z^{-1}(\tilde{\Omega}^{-1}\tilde{B} + X'Y). \tag{A1}$$

The derivative of $\overline{B}$ with respect to λ can be computed from:

$$\begin{aligned}\frac{\partial \overline{B}}{\partial \tilde{\lambda}} &= Z^{-1}\left(\frac{\partial}{\partial \tilde{\lambda}}(\tilde{\Omega}^{-1}\tilde{B} + X'Y)\right) + \left(\frac{\partial Z^{-1}}{\partial \tilde{\lambda}}\right)(\tilde{\Omega}^{-1}\tilde{B} + X'Y), \\ &= Z^{-1}\left(\frac{\partial}{\partial \tilde{\lambda}}\tilde{\Omega}^{-1}\tilde{B}\right) + \left(\frac{\partial Z^{-1}}{\partial \tilde{\lambda}}\right)(\tilde{\Omega}^{-1}\tilde{B} + X'Y).\end{aligned} \tag{A2}$$

From Magnus and Neudecker (1999), the derivative of an inverse matrix of functions Z can be written as:

$$\frac{\partial Z^{-1}}{\partial \tilde{\lambda}} = -Z^{-1}\left(\frac{\partial Z}{\partial \tilde{\lambda}}\right)Z^{-1}. \tag{A3}$$

Since the matrix $X'X$ is not a function of $\tilde{\lambda}$, the derivative $\partial Z / \partial \tilde{\lambda}$ is:

$$\frac{\partial Z}{\partial \tilde{\lambda}} = \frac{\partial \tilde{\Omega}^{-1}}{\partial \tilde{\lambda}} = -\frac{2}{\tilde{\lambda}^3} diag(\sigma_1^2,...,\sigma_m^2; 2^2 \cdot \sigma_1^2,...,2^2 \cdot \sigma_m^2;...; p^2 \cdot \sigma_1^2,..., p^2 \cdot \sigma_m^2; 0). \tag{A4}$$

The derivative $\partial(\tilde{\Omega}^{-1}\tilde{B}) / \partial \tilde{\lambda}$ can also be written as:

$$\frac{\partial}{\partial\tilde{\lambda}}\tilde{\Omega}^{-1}\tilde{B}=\left(\frac{\partial\tilde{\Omega}^{-1}}{\partial\tilde{\lambda}}\right)\tilde{B}, \tag{A5}$$

and the value of $\partial\tilde{\Omega}^{-1}/\partial\tilde{\lambda}$ is as in (A4).

Totally, from (A1) – (A5), we have:

$$\begin{aligned}\frac{\partial\bar{B}}{\partial\tilde{\lambda}}&=Z^{-1}\left(\frac{\partial\tilde{\Omega}^{-1}}{\partial\tilde{\lambda}}\right)\tilde{B}-Z^{-1}\left(\frac{\partial\tilde{\Omega}^{-1}}{\partial\tilde{\lambda}}\right)Z^{-1}(\tilde{\Omega}^{-1}\tilde{B}+X'Y),\\&=Z^{-1}\left(\frac{\partial\tilde{\Omega}^{-1}}{\partial\tilde{\lambda}}\right)(\tilde{B}-\bar{B}).\end{aligned} \tag{A6}$$