JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **30**: 46–73 (2015)
Published online 26 March 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2315

BAYESIAN VARS: SPECIFICATION CHOICES AND FORECAST ACCURACY

ANDREA CARRIERO[a]*, TODD E. CLARK[b] AND MASSIMILIANO MARCELLINO[c,d,e]

[a] *Queen Mary, University of London, UK*
[b] *Federal Reserve Bank of Cleveland, OH, USA*
[c] *European University Institute, Florence, Italy*
[d] *Bocconi University, Milan, Italy*
[e] *CEPR, London, UK*

SUMMARY

In this paper we discuss how the point and density forecasting performance of Bayesian vector autoregressions (BVARs) is affected by a number of specification choices. We adopt as a benchmark a common specification in the literature, a BVAR with variables entering in levels and a prior modeled along the lines of Sims and Zha (*International Economic Review* 1998; **39**: 949–968). We then consider optimal choice of the tightness, of the lag length and of both; evaluate the relative merits of modeling in levels or growth rates; compare alternative approaches to *h*-step-ahead forecasting (direct, iterated and pseudo-iterated); discuss the treatment of the error variance and of cross-variable shrinkage; and assess rolling versus recursive estimation. Finally, we analyze the robustness of the results to the VAR size and composition (using also data for France, Canada and the UK, while the main analysis is for the USA). We obtain a large set of empirical results, but the overall message is that we find very small losses (and sometimes even gains) from the adoption of specification choices that make BVAR modeling quick and easy, in particular for point forecasting. This finding could therefore further enhance the diffusion of the BVAR as an econometric tool for a vast range of applications. Copyright © 2013 John Wiley & Sons, Ltd.

Received 25 August 2011; Revised 15 November 2012

Supporting information may be found in the online version of this article.

1. INTRODUCTION

Forecasting future developments in the economy is a key element of the decision process in policy making, consumption and investment decisions, and financial planning. For example, members of the Federal Open Market Committee often stress that, because monetary policy affects the economy with a lag, policy must be forward-looking. Looking ahead means relying on forecasts of output growth, inflation, and other key indicators.

Recently there has been a resurgence of interest in applying Bayesian methods to point and density forecasting, particularly with Bayesian vector autoregressions (BVARs). BVARs have a long history in forecasting, stimulated by their effectiveness documented in the seminal studies of Doan *et al*. (1984) and Litterman (1986). In recent years, the models seem to be used even more systematically for policy analysis and forecasting macroeconomic variables (e.g. Kadiyala and Karlsson, 1997; Koop, 2013). At present, there is considerable interest in using BVARs for these purposes in a large dataset context (e.g. Carriero *et al*., 2009, 2011; Banbura *et al*., 2010; Koop, 2013).

However, putting BVARs to use in practical forecasting raises a host of detailed questions about model specification, estimation and forecast construction.

With regard to model specification, the researcher needs to address issues such as (i) the choice of the tightness and of the lag length of the BVAR; (ii) the treatment of the error variance and the

* Correspondence to: Andrea Carriero, Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS, UK. E-mail: a.carriero@qmul.ac.uk

imposition of cross-variable shrinkage[1]; (iii) whether or not to transform the variables to get stationarity, and whether to complement this choice with the imposition of priors favoring cointegration and unit roots.

Accordingly, a first point of this paper is to examine the effects that such specification choices have on the forecast accuracy of the BVAR. We adopt as a benchmark a common specification in the literature, a Bayesian VAR with variables entering in levels and a prior modeled along the lines of Sims and Zha (1998), Robertson and Tallman (1999), Waggoner and Zha (1999), Zha (1998) and, more recently, Giannone *et al*. (2012).

With regard to model estimation and forecast construction, under some approaches estimating and forecasting with a BVAR can be technically and computationally demanding. For the homoskedastic BVAR with natural conjugate prior, the posterior and one-step-ahead predictive densities have convenient analytical forms (Student's *t*). However, even for this prior, multi-step predictive densities do not have analytical forms and simulation methods are required. Under a Normal-inverted Wishart prior and posterior that treat each equation symmetrically, Monte Carlo methods can be used to efficiently simulate the multi-step predictive densities, taking advantage of a Kronecker structure to the posterior variance of the model's coefficients.

Other priors or model extensions (such as allowing for asymmetric prior variances across equations, or time series heteroskedasticity in the disturbances) mean that neither posteriors nor predictive densities have analytical forms. In these cases, simulations become more computationally intensive because the posterior variance of the model's coefficients no longer has a Kronecker structure. To avoid costly simulation, Litterman's (1986) specification of the Minnesota prior treats the error variance matrix as fixed and diagonal. Litterman (1986) imposes such a strong assumption to allow for equation-by-equation ridge estimation of the system; treating the error variance matrix as random would have required Markov chain Monte Carlo (MCMC) simulations of the entire system of equations.

While improved computational power has made simulation of models under a Normal-inverted Wishart prior specification more tractable, some researchers and practitioners may prefer to avoid simulation methods and use alternatives considered in such studies as Banbura *et al*. (2010) and Koop (2013). This preference could stem from the computational hurdles of conducting simulations with very large models. Also, it can stem from very tight time constraints for the production of the forecasts, as can be the case for market strategists, for example. Alternatively, the preference could be a function of software choice and the coding burden of simulation. Common software such as RATS and Eviews provides commands for estimating BVARs and forecasting without simulation; simulation requires more significant programming by the user. Similarly, while many users of Matlab are capable of programming simulation, the absence of simple procedures or toolboxes may make simulation costly to other users.

Accordingly, a second point of this paper is to examine approaches that make the computation of point and density forecasts from BVARs quick and easy, for example by making specific choices on the priors and by using direct rather than iterated forecasts (e.g. Marcellino *et al*., 2006). In most cases, the resulting forecasts represent approximations of the posterior distribution. Hence we then assess whether such approximations yield significant losses in terms of decreased forecast precision, as measured by either the root mean squared forecast error or the predictive score, in the case of density forecasts. We show that, for users focused on point forecasts, there is little cost to methods that do not involve simulation.

Since it is difficult to rank the alternative modeling and forecasting choices from a purely theoretical point of view, given that their relative performance will be determined by the unknown data-generating process, we take a more practical perspective. Specifically, we consider a set of variables whose future evolution is of key interest for central banks and more generally for economic policy making, and we

[1] That is, whether to shrink more towards 0 the coefficients related to the regressors which are not lags of the dependent variable of a given equation.

DOI: 10.1002/jae

evaluate the performance of different BVAR modeling choices in this context. In light of recent evidence on the success of larger models relative to smaller ones and interest in large datasets (e.g. Banbura *et al*., 2010; Koop, 2013), we focus on mid-size models applied to monthly data: 18-variable BVARs for US macroeconomic and financial data.

To ensure our results have broad applicability, we check their robustness to changes in both the time series and cross-sectional dimension of the system. In particular, we consider recursive and rolling estimation, a reduction in the size of the VAR to a subset of seven of the 18 US variables, and we repeat the analysis for some other datasets—specifically, data for Canada, France and the UK.

We obtain a large set of empirical results, but we can summarize them by saying that we find very small losses (and sometimes even gains) from the adoption of BVAR modeling choices that make forecast computation quick and easy, in particular for point forecasting. An approach that works well is to specify a Normal-inverted Wishart prior along the lines of Sims and Zha (1998) on the VAR in levels, preferably optimizing its tightness and lag length. Optimizing over the lag length is generally helpful, and optimal selection of the tightness never harms, though the average gains are small in our empirical applications. For the accuracy of point forecasts, there proves to be essentially no payoff to using MCMC methods to obtain multi-step forecasts from the posterior distribution. For density forecasting, simulation methods work better than a direct multi-step approach, especially at longer horizons (and less so at shorter horizons). Specifications in levels benefit a lot from the imposition of the sum of coefficients and dummy initial observation priors of Doan *et al*. (1984) and Sims (1993). Instead, there is no payoff to using a Litterman (1986) prior that treats the error variance matrix as fixed and diagonal and is tighter for lags of other variables than for lags of the dependent variable. Using forecast robustifying methods, such as rolling estimation or modeling in differences, can enhance the density forecasting performance, while in terms of mean squared error it is difficult to do better than the benchmark. The finding that simple methods work well could therefore further enhance the diffusion of the BVAR as an econometric tool for a vast range of applications.

The paper is structured as follows. In Section 2 we describe the US data and the design of the forecasting exercise. In Section 3 we present the baseline case. In the following three sections we evaluate changes in three main features of the benchmark. Specifically, In Section 4 we consider optimal choice of the tightness, lag length and both. In Section 5 we consider modeling in levels or growth rates. In Section 6 we compare the alternative approaches to multi-step forecasting, with a special focus on the non-simulation-based ones. In Section 7 we discuss the treatment of the error variance and of cross-variable shrinkage. Next we evaluate the robustness of our findings. In particular, in Section 8 we consider the relative merits of rolling and recursive estimation. In Section 9 we look at the size of the VAR and in Section 10 we summarize the results for Canada, France and the UK, comparing them with those for the US. Finally, in Section 11 we summarize the main findings and conclude. Supplemental material referred to in the text is available upon request.

2. DATA AND DESIGN OF THE FORECASTING EXERCISE

Our dataset for the USA has monthly frequency and runs from January 1973 to March 2010.[2] The data include 18 macroeconomic and financial series of major interest to policymakers and forecasters, listed in Table I (panel A).

In the paper we will report results based on both a VAR for the variables in levels or log-levels (which we label VAR in levels), and a VAR estimated after transforming variables as needed to get stationarity (which we label VAR in growth rates). In this growth rates specification, we log-difference

[2] While most of the US data series are available before 1973, the exchange rate index is only available back to 1973. Like most other studies assessing forecasting with larger models, we do not consider real-time data, owing to limited availability for the series of interest.

DOI: 10.1002/jae

Table I. Description of dataset and transformations

Code	Series	Transformation	
		VAR in Levels	VAR in growth rates
Panel A: *USA*			
UR	Unemployment rate	None	None
PCEPI	PCE price index	$1200ln y_t$	$1200ln(y_t/y_{t-1})$
PCEXFEPI	Core PCE price index (ex food and energy)	$1200ln y_t$	$1200ln(y_t/y_{t-1})$
PAYROLLS	Nonfarm payroll employment	$1200ln y_t$	$1200ln(y_t/y_{t-1})$
WEEKLYHRS	Weekly hours worked	None	None
CLAIMS	New claims for unemployment insurance	None	None
RETAILSALES	Nominal retail sales	$1200ln y_t$	$1200ln(y_t/y_{t-1})$
CONSCONF	Index of consumer confidence	None	None
STARTS	Single-family housing starts	$100ln y_t$	$100ln(y_t/y_{t-1})$
IP	Industrial production	$1200ln y_t$	$1200ln(y_t/y_{t-1})$
CU	Index of capacity utilization	None	None
PMISUPDELIV	Purchasing Managers' Index of supplier delivery times	None	None
PMIORDERS	Purchasing Managers' Index of new orders	None	None
POIL	Price of oil (West Texas Intermediate)	$100ln y_t$	$100ln(y_t/y_{t-1})$
SP500	S&P 500 index of stock prices	$100ln y_t$	$100ln(y_t/y_{t-1})$
ITB10y	Yield on 10-year Treasury bonds	None	None
FFR	Federal funds rate	None	None
REALXR	Real exchange rate	$100ln y_t$	$100ln(y_t/y_{t-1})$
Panel B: *Canada*, *France*, *UK*			
UNRATE	Unemployment rate	None	None
EMPLOY	Total employment	$1200ln(y_t)$	$1200ln(y_t/y_{t-1})$
IP	Industrial production	$1200ln(y_t)$	$1200ln(y_t/y_{t-1})$
CPI	CPI inflation	$1200ln(y_t)$	$1200ln(y_t/y_{t-1})$
OIL	Spot commodity price—crude oil	$100ln(y_t)$	$100ln(y_t/y_{t-1})$
XRATE	Real exchange rate vs. major currencies	$100ln(y_t)$	$100ln(y_t/y_{t-1})$
STOCKPRICE	Stock price index	$100ln(y_t)$	$100ln(y_t/y_{t-1})$
POLRATE	Official policy rate	None	None
BONDRATE	10-year government bond yield	None	None

Note: The used stock price index is TSE-300 for Canada, SPF-250 for France, and FTSE-100 for the UK. The used policy rate is Overnight target rate for Canada, Banque de France Official Lending Rate and ECB policy rate for France, and Bank of England official bank rate for the UK. Data are taken from the Forecasting Analysis and Modeling Environment Database, OECD, Conference board, BIS, ECB and Bank of England.

variables such as employment to make them stationary, but we do not difference interest rates and diffusion indexes from surveys because, conceptually, they should be stationary. For all variables, the prior means of the coefficients will be set accordingly to 1 (for the VAR in levels) or 0 (for the VAR in growth rates). The transformations used on each variable are listed in Table I.

While we estimate the models in both levels form and (for some variables) difference form, we always report the forecast results in units corresponding to stationary variables, given in the last column of Table I. For example, the forecast results for industrial production (IP) are for annualized growth rates of IP. For models estimated in levels, we must transform some of the model-produced forecasts to use the same units.

The main forecasting exercise is performed in pseudo-real time, i.e. we never use information which is not available at the time the forecast is made. For all models, we use a recursive estimation window, except in section 8, where we assess the robustness of the results to the use of a rolling sample estimation scheme. We have data starting from 1973:1, but after differencing the first observation is missing. Moreover, as we plan to compare models in levels featuring up to 13 lags (and 12 lags in growth rates), we start with the estimation sample of 1974:2 to 1985:12 in order to have the same number of data points for each model. We produce forecasts for all the horizons up to 12 steps ahead; for a horizon of h periods, the first available forecast is for 1986:1 + *h-1*. Our last estimation sample is 1974:2 to 2009:3, yielding a forecast for horizon h for date 2009:4 + *h-1*.

DOI: 10.1002/jae

We will evaluate both the point and density forecast ability of the examined models. For point forecasts, we evaluate our results in terms of root mean squared forecast error (RMSFE) for a given model. Let $\hat{y}^{(i)}_{t+h}(M)$ denote the forecast of the ith variable $y^{(i)}_{t+h}$ made by model M. The RMSFE made by model M in forecasting the ith variable at horizon h is

$$\text{RMSFE}^{M}_{i,h} = \sqrt{\frac{1}{P}\sum\left(\hat{y}^{(i)}_{t+h}(M) - y^{(i)}_{t+h}\right)^2} \tag{1}$$

where the sum is computed over all the P forecasts produced.

The overall calibration of the density forecasts can be measured with log predictive density scores, motivated and described in Geweke and Amisano (2010), for example.[3] At each forecast origin, we compute the log predictive score using the quadratic approximation of Adolfson *et al*. (2007). Specifically, using the simulated distributions of forecasts, we compute the log score in predicting variable i as

$$s_t\left(y^{(i)}_{t+h}\right) = -0.5\left[ln(2\pi) + ln\left(V^{i}_{t+h|t}\right) + \left(y^{(i)}_{t+h} - \bar{y}_{t+h|t}\right)^2 / V^{i}_{t+h|t}\right] \tag{2}$$

where $\bar{y}_{t+h|t^i}$ and $V_{t+h|t^i}$ denote the posterior mean and variance of the simulated forecast distribution for variable i at h steps ahead. The average score obtained by model M in predicting variable i h steps ahead is

$$\text{SCORE}^{M}_{i,h} = \frac{1}{P}\sum s_t\left(y^{(i)}_{t+h}\right) \tag{3}$$

To compare each model M against the benchmark B we therefore consider the percentage gains in terms of RMSFE, defined as

$$\left(1 - \text{RMSFE}^{M}_{i,h}/\text{RMSFE}^{B}_{i,h}\right) \times 100 \tag{4}$$

and the percentage gain in terms of score, which is

$$\left(\text{SCORE}^{M}_{i,h} - \text{SCORE}^{B}_{i,h}\right) \times 100 \tag{5}$$

Finally, to have an indication of the statistical significance of differences in forecasting performance, we provide the results of the Diebold and Mariano (1995) test for equal mean squared forecast error, compared against standard normal critical values. Following the recommendation of Clark and McCracken (2011c), to reduce the chances of spurious rejections at longer forecast horizons, we compute the Diebold–Mariano test with the Harvey *et al*. (1997) adjustment of the variance that enters the test statistic. Our use of the Diebold–Mariano test with forecasts that are, in many cases, nested is a deliberate choice. Monte Carlo evidence in Clark and McCracken (2011a,2011b) indicates that, with nested models, the Diebold–Mariano test compared against normal critical values can be viewed as a somewhat conservative (in the sense of tending to have size modestly below nominal size) test for equal accuracy in finite samples. Nonetheless, we obtain many rejections of the null of equal accuracy. This reflects the higher power of the test due to our long forecast sample (269 one-step observations for the USA) compared to many other studies in the literature.

To provide a rough gauge of the statistical significance of differences in average log scores, we use the Amisano and Giacomini (2007) t-test of equal means, applied to the log score for each model relative to the baseline forecast. We view the tests as a rough gauge because the asymptotic validity of the Amisano and Giacomini (2007) test requires that, as forecasting moves forward in time, the

[3] The scores are averages of predictive likelihoods. The predictive likelihood is closely related to the marginal likelihood, as the marginal likelihood can be expressed as the product of a sequence of one-step-ahead predictive likelihoods.

DOI: 10.1002/jae

models be estimated with a rolling, rather than expanding, sample of data. The *t*-statistics are computed with a serial correlation-robust variance, using a rectangular kernel, *h-1* lags, and the small-sample adjustment of Harvey *et al.* (1997).

3. BASELINE CASE

3.1. Baseline Specification

The baseline specification, against which we will compare alternative modeling choices, is a standard BVAR with a Normal-inverted Wishart (N-IW) conjugate prior. Given N different variables grouped in the vector $y_t = (y_{1t}\, y_{2t} \ldots y_{Nt})'$, we consider the following VAR:

$$y_t = \Phi_c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \ldots + \Phi_p y_{t-p} + \varepsilon_t; \varepsilon_t \sim i.i.d.N(0, \Sigma) \tag{6}$$

where $t = 1, \ldots, T$. Each equation has $M = Np + 1$ regressors. By grouping the coefficient matrices in the $N \times M$ matrix $\Phi = \left[\Phi_c\, \Phi_1 \ldots \Phi_p\right]$ and defining $x_t = (1\ y'_{t-1} \ldots y'_{t-p})'$ as a vector containing an intercept and p lags of y_t, the VAR can be written as

$$y_t = \Phi' x_t + \varepsilon_t. \tag{7}$$

An even more compact notation is

$$Y = X\Phi + E \tag{8}$$

where $Y = [y_1, \ldots, y_T]'$, $X = [x_1, \ldots, x_T]'$, and $E = [\varepsilon_1, \ldots, \varepsilon_T]'$ are, respectively, $T \times N$, $T \times M$ and $T \times N$ matrices. Finally, for representing multi-step forecasts, another useful notation is the companion form:

$$x_{t+1} = \Phi^+ x_t + \widetilde{\varepsilon}_t \tag{9}$$

where $\widetilde{\varepsilon}_t$ is an $M \times 1$ vector containing ε_t and 0's elsewhere and

$$\Phi^+ = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \Phi_c & \Phi_1 & \Phi_2 & \ldots & \Phi_p \\ 0 & I_N & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ 0 & 0 & & I_N & 0 \end{bmatrix} \tag{10}$$

Note that in this notation y_t corresponds to rows $2, \ldots, N+1$ of x_{t+1}, so we can write $y_t = \mathbf{s}x_{t+1}$, defining $\mathbf{s}$ to be a selection matrix selecting the appropriate rows (i.e. row 2 to row $N+1$) of x_{t+1}. With this representation, multi-step forecasts can be obtained as $\hat{x}_{t+h} = (\Phi^+)^h x_t$.

We use the conjugate N-IW prior:

$$\Phi|\Sigma \sim N(\Phi_0, \Sigma \otimes \Omega_0), \Sigma \sim \mathrm{IW}(S_0, v_0) \tag{11}$$

As the N-IW prior is conjugate, the conditional posterior distribution of this model is also N-IW (Zellner 1971):

$$\Phi|\Sigma, Y \sim N(\bar{\Phi}, \Sigma \otimes \bar{\Omega}), \Sigma|Y \sim \mathrm{IW}(\bar{S}, \bar{v}) \tag{12}$$

Defining $\hat{\Phi}$ and $\hat{E}$ as the OLS estimates, we have that $\bar{\Phi} = \left(\Omega_0^{-1} + X'X\right)^{-1}\left(\Omega_0^{-1}\Phi_0 + X'Y\right)$, $\bar{\Omega} = \left(\Omega_0^{-1} + X'X\right)^{-1}$, $\bar{v} = v_0 + T$, and $\bar{S} = \Phi_0 + \hat{E}'\hat{E} + \hat{\Phi}'X'X\hat{\Phi} + \Phi'_0\Omega_0^{-1}\Phi_0 - \bar{\Phi}'\bar{\Omega}^{-1}\bar{\Phi}$. In the

DOI: 10.1002/jae

case of the natural conjugate N-IW prior, the marginal posterior distribution of Φ is matricvariate-t with expected value $\bar{\Phi}$.

Finally, in the baseline specification in levels we choose a lag length of 13.

3.2. Baseline Prior Parametrization

In our baseline specification we impose the prior expectation and standard deviation of the coefficient matrices to be

$$E\left[\Phi_k^{(ij)}\right] = \begin{cases} \Phi^* \ \text{if} i = j, \ \ k = 1 \\ 0 \ \text{otherwise} \end{cases}, \text{SD}\left[\Phi_k^{(ij)}\right] = \begin{cases} \dfrac{\lambda_1 \lambda_2}{k} \dfrac{\sigma_i}{\sigma_j}, \ \ k = 1, \ldots, p \\ \lambda_0 \sigma_i, \ \ k = 0 \end{cases} \tag{13}$$

where $\Phi_k^{(ij)}$ denotes the element in position (i,j) in the matrix Φ_k. The prior mean Φ^* is set to 1 in the VAR in levels specifications and to 0 in the VAR in growth rates specification. For the intercept we assume an informative prior with mean 0 and standard deviation $\lambda_0 \sigma_i$. The shrinkage parameter λ_1 measures the overall tightness of the prior: when $\lambda_1 \to 0$ the prior is imposed exactly and the data do not influence the estimates, while as $\lambda_1 \to \infty$ the prior becomes loose and the prior information does not influence the estimates, which will approach the standard OLS estimates. The parameter λ_2 implements additional shrinkage on lags of other variables than for lags of the dependent variable. We refer to this as the cross-shrinkage parameter, and in our baseline specification we set it to $\lambda_2 = 1$, which implies that no cross-variable shrinkage takes place, as required for the Normal-inverted Wishart case. To set each scale parameter σ_i we follow common practice (see, for example, Litterman, 1986; Sims and Zha, 1998) and set it equal to the standard deviation of the residuals from a univariate autoregressive model.

Note that the prior beliefs in (13), defining the traditional Minnesota prior, only include the prior mean and variances of the coefficients, and do not elicit any prior beliefs about the correlations among the coefficients. Doan *et al*. (1984) and Sims (1993) have proposed to complement the prior beliefs in (13) with additional priors which favor unit roots and cointegration, and introduce correlations in prior beliefs about the coefficients in a given equation. Both these priors were motivated by the need to avoid having an unreasonably large share of the sample period variation in the data accounted for by deterministic components (Sims, 1993). These priors are also in line with the belief that macroeconomic data typically feature unit roots and cointegration.

Accordingly, in our benchmark specification, to the prior beliefs in (13) we add the 'sum of coefficients' and 'dummy initial observation' priors proposed in Doan *et al*. (1984) and Sims (1993), respectively. Both these priors can be implemented by augmenting the system with dummy observations.

More specifically, the 'sum of coefficients' prior expresses a belief that when the average of lagged values of a variable is at some level $\bar{y}_{0i}$, that same value $\bar{y}_{0i}$ is likely to be a good forecast of future observations, and is implemented by augmenting the system in (8) with the dummy observations Y_{d_1} and X_{d_1} with generic elements:

$$y_d(i,j) = \begin{cases} \bar{y}_{0i}/\lambda_3 \ \ \text{if} i = j \\ 0 \ \ \text{otherwise} \end{cases}; x_d(i,s) = \begin{cases} \bar{y}_{0i}/\lambda_3 \ \ \text{if} i = j, \ \ s < M \\ 0 \ \ \text{otherwise} \end{cases} \tag{14}$$

where i and j go from 1 to N while s goes from 1 to M. When $\lambda_3 \to 0$ the model tends to a form that can be expressed entirely in terms of differenced data, there are as many unit roots as variables and there is no cointegration.

The 'dummy initial observation' prior introduces a single dummy observation such that all values of all variables are set equal to the corresponding averages of initial conditions up to a scaling factor $(1/\lambda_4)$. It is implemented by adding to the system in (8) the dummy variables Y_{d_2} and X_{d_2} with generic elements:

DOI: 10.1002/jae

$$y_d(j) = \bar{y}_{0j}/\lambda_4; x_d(s) = \begin{cases} \bar{y}_{0j}/\lambda_4 & \text{for} s < M \\ 1/\lambda_4 & \text{for} s = M \end{cases} \tag{15}$$

where j goes from 1 to N while s goes from 1 to M. As $\lambda_4 \to 0$ the model tends to a form in which either all variables are stationary with means equal to the sample averages of the initial conditions, or there are unit root components without drift terms, which is consistent with cointegration.

To summarize, the prior mean and variance of the VAR coefficients in our benchmark specification are represented by combining a fictitious sample that imposes the mean and variances in (13) (in the interest of brevity, we refer the reader to sources such as Banbura *et al*. (2010) and Karlsson (2012) for the details in lieu of spelling out here the necessary dummy variable matrices) with the fictitious samples given by the dummy variables defined by (14) and (15). The prior specification is completed by choosing v_0 and S_0 so that the prior expectation of Σ is equal to a fixed diagonal residual variance $E[\Sigma] = \text{diag}\left(\sigma_1^2, \ldots, \sigma_N^2\right)$. In particular, following Kadiyala and Karlsson (1997), we set the diagonal elements of S_0 to $s_{0ii} = (v_0 - N - 1)\sigma_i^2$ and $v_0 = N + 2$.

This prior is similar to that proposed by Sims and Zha (1998), with the subtle difference that in the original implementation the prior is elicited on the coefficients of the structural representation of the VAR rather than on the reduced form.[4] This prior has been widely used in the literature (see, for example, Leeper *et al*., 1996; Sims and Zha, 1996; Zha, 1998; Robertson and Tallman, 1999; Waggoner and Zha, 1999; and more recently, Giannone *et al*., 2012), and it is available ready-packaged in at least two widely used econometric software packages (Eviews and RATS). Finally, we note the baseline prior is very close to that used by Banbura *et al*. (2010), the only difference being the addition of the 'dummy initial observation' component prior.[5]

Most of the studies cited above have considered the following parametrization for the prior, which we adopt in our baseline model specification:

$$\lambda_0 = 1; \lambda_1 = 0.2; \lambda_2 = 1; \lambda_3 = 1; \lambda_4 = 1. \tag{16}$$

After discussing the results for this baseline specification, in the following sections we will assess the consequences of varying λ_1 (the tightness), λ_2 (related to cross-variable shrinkage), λ_3, λ_4 (related, respectively, to the unit root and cointegration prior), and modeling in growth rates rather than levels.

We note here that our baseline model is featuring the N-IW natural conjugate prior, which has several advantages and some costs. In particular, as is clear from equation (11), the prior variance of the coefficients has a Kronecker structure. This means that the prior variances of the coefficients are symmetric across equations (they differ only up to a scaling factor given by the elements of the error variance Σ). This structure of the prior variance matrix is what gives rise to the conjugacy of the prior, as it implies that also the posterior in (12) has the same form. As we shall see, this leads to relevant computational gains. However, the assumption that the prior variances of the coefficients are symmetric across equations can be restrictive. For example, in the original Litterman (1986) implementation of the prior, the parameter λ_2 is fixed to a value smaller than 1, reflecting the belief that own-lags are more relevant than lags of other variables in explaining the behavior of a given dependent variable in a given equation of the VAR. While such a belief makes sense, it is not implementable within the N-IW conjugate framework, as fixing λ_2 to a value different from 1 would break down the Kronecker structure in (11) and therefore the conjugacy of the prior and posterior.

[4] Note that this prior does not necessarily have the same N-IW conjugate form as that in Kadiyala and Karlsson (1997), but it is still computationally tractable. As discussed by Sims and Zha (1998), this prior is more general, and will coincide with that in Kadiyala and Karlsson (1997) when the prior covariance matrix in the structural representation of the VAR is the same across equations. This is the case in our forecasting application.

[5] Our specifications also differ in that Banbura *et al*. (2010) use prior means of 0 on some first-lag coefficients, while we follow studies such as Sims and Zha (1998) in using prior means of 1 on all first-lag coefficients in our baseline specification in levels.

DOI: 10.1002/jae

3.3. Forecasting

Under the standard N-IW prior described above, the full distribution of the one-step-ahead forecasts is matricvariate-t (MT):

$$y'_{T+1}|x'_{T+1} \sim MT\left(x'_{T+1}\bar{\Phi}, \left(x'_{T+1}\bar{\Omega}x_{T+1}+1\right)^{-1}, \bar{S}, \bar{\nu}\right) \tag{17}$$

Multi-step-ahead forecasts obtained by iteration are not available in closed form, but can be simulated using an MC algorithm which draws a sequence of Σ and parameters Φ from (12) and shocks and at each draw j computes the implied path of $\hat{y}^{(j)}_{t+h}$. Drawing a sequence of Φ can, in general, be rather demanding from a computational point of view, but in this specific case the matricvariate structure of the N-IW prior ensures there are efficient algorithms that considerably speed up the computations. An intuitive way to draw Φ, conditionally on a draw of the error variance Σ, is to vectorize it and draw from a multivariate normal. In this case a draw of Φ from (12) is obtained as follows:

$$\text{vec}(\Phi) = \text{vec}(\bar{\Phi}) + \text{chol}(\Sigma \otimes \bar{\Omega}) \times v \tag{18}$$

where v is an $MN \times 1$ standard Gaussian vector process. The Choleski decomposition above requires $(MN)^3$ elementary operations. The scheme outlined in (18) does not take advantage of the matricvariate structure of the distribution of Φ. Indeed, by organizing the elements of v in an $M \times N$ matrix V such that $v = \text{vec}(V)$, one could draw the matrix Φ as follows:

$$\Phi = \bar{\Phi} + \text{chol}(\bar{\Omega}) \times V \times \text{chol}(\Sigma)' \tag{19}$$

This can considerably speed up the computations, because the two Choleski decompositions $\text{chol}(\bar{\Omega})$ and $\text{chol}(\Sigma)$ require only $M^3 + N^3$ operations, but can only be implemented when the variance matrix of the prior coefficients has a Kronecker structure.

Table II in the supporting information reports, for each of the 18 variables under evaluation, the RMSFEs (panel A) and average log scores (panel B) over the entire forecast sample 1986–2010. Most of the subsequent results will be expressed as values relative to these baseline losses.

Note that, as mentioned, following Banbura *et al*. (2010), among others, we estimate the baseline BVAR in levels but produce and evaluate forecasts on the growth rates for the trending variables identified in Table I, as these are the quantities of interest for such a prototypical macroeconomic dataset.

4. SELECTION OF HYPERPARAMETERS AND LAG LENGTH

4.1. Selection of Hyperparameters (Tightness)

To make the prior operational, one needs to choose the value of the hyperparameter λ_1 which controls the overall tightness of the prior. To consider the effect of prior optimization on forecast accuracy, we follow Carriero *et al*. (2012) and at each point we choose θ by maximizing the marginal data density of the model[6]:

$$\lambda^*_{1t} = arg\max_{\lambda_1} lnp(Y) \tag{20}$$

[6] Giannone *et al*. (2012) propose a similar strategy for forecasting a macroeconomic dataset. While our strategy implicitly assumes a flat prior on a discrete set of possible values for λ_1, their strategy assumes a proper (albeit uninformative) prior on a continuum of values. Still other studies, such as Banbura *et al*. (2010), have selected hyperparameters using alternative strategies based on RMSFEs.

DOI: 10.1002/jae

The marginal data density (or marginal likelihood) $p(Y)$ can be obtained by integrating out all the coefficients of the model. Defining Θ as the set of all the coefficients of the model, we have

$$p(Y) = \int p(Y|\Theta)p(\Theta)d\Theta \tag{21}$$

Under our Normal-inverted Wishart prior the density $p(Y)$ can be computed in closed form (Zellner, 1971; Bauwens *et al*., 1999) and is given by

$$p(Y) = \pi^{\frac{-TN}{2}} \times \left|\left(I + X\Omega_0 X'\right)^{-1}\right|^{\frac{N}{2}} \times \left|S_0\right|^{\frac{v_0}{2}} \times \frac{\Gamma_N\left(\frac{v_0+T}{2}\right)}{\Gamma_N\left(\frac{v_0}{2}\right)}$$

$$\times \left|S_0 + \left(Y - X\Phi_0\right)'\left(I + X\Omega_0 X'\right)^{-1}(Y - X\Phi_0)\right|^{-\frac{v_0+T}{2}} \tag{22}$$

with $\Gamma_N(\cdot)$ denoting the N-variate gamma function. A straightforward derivation based on theorem A.19 in Bauwens *et al*. (1999) can be found in Carriero *et al*. (2012). The value of the marginal likelihood in (22) is provided by default in some computer packages such as Eviews.

We now present the results obtained when in each time period we set $\lambda_1 = \lambda_{1t}^*$, i.e. we set the tightness to the value maximizing the marginal likelihood. We optimize over a discrete grid $\lambda_1 \in \{$ 0.01, 0.025, 0.050, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1, 2, 5 $\}$. For our baseline model, it turns out that the value of the tightness does not change over time and remains at the value 0.15. As in the baseline specification the shrinkage is fixed at 0.2, we do not expect great losses or gains from using the optimal tightness.

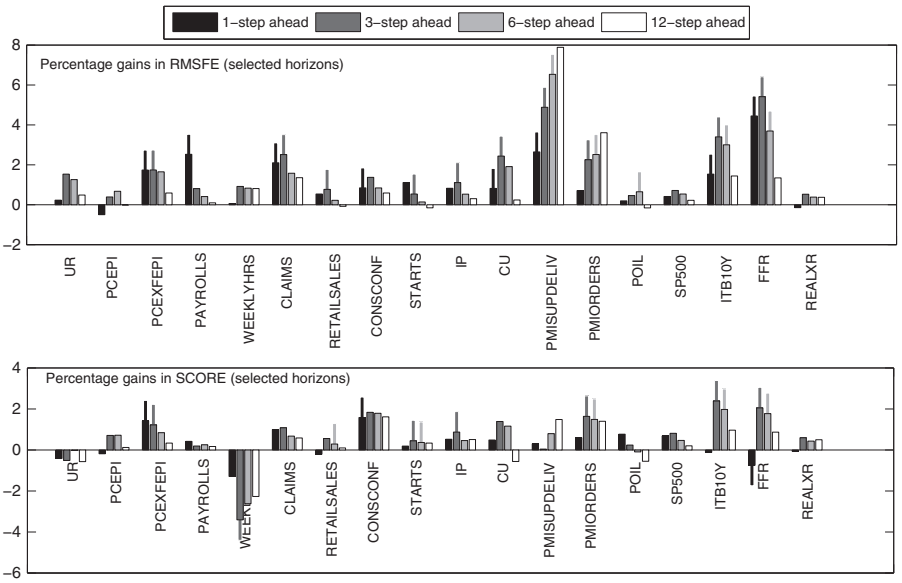
Figure 1 shows the gains in RMSFE and predictive SCORE obtained by using $\lambda_1 = \lambda_{1t}^*$ (with a fixed lag length of 13) with respect to the baseline model (which fixes the lag order at 13 and the tightness hyperparameter at $\lambda_1 = 0.2$). Note that, to facilitate presentation, the chart in the upper panel focuses on horizons of 1, 3, 6 and 12 periods; the summary table in the lower panel covers all horizons between 1 and 12 periods. As is clear, optimizing the tightness rarely yields losses. However, as λ_{1t}^* turns out to be very close to the baseline value, the gains are rather small—on average 1.41% over all variables and horizons—though often statistically significant at shorter horizons. The RMSFE gains are higher for some variables such as the Purchasing Managers' Index of supplier delivery times, the 10-year yield and the federal funds rate (FFR). A similar picture emerges when looking at density forecasting: the gains are overall positive but rather small, with an average/median gain of about 0.5%.

We have also considered choosing optimally the hyperparameters λ_3 and λ_4 (related respectively to the 'sum of coefficients' and 'initial dummy observation' prior) by optimizing the marginal likelihood with respect to them. For λ_1 we used a grid of values $\{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5\}$, while for λ_3 and λ_4 we used the grid $\{0.25, 0.5, 1, 1.5, 2, 2.5, 3\}$. Results for this strategy are not reported, for brevity, but can be found in the supporting information, Figure 10. While the resulting forecasts are slightly superior to the benchmark on average, such a strategy yields slightly smaller gains on average with respect to the case where λ_1, λ_3 and λ_4 are kept fixed and only the lag length is optimized (see the discussion in the next subsection, and Figure 2).

4.2. Selection of the Lag Length

Up to now we have assumed a fixed lag length, p=13. However, the researcher needs to determine the lag length as well. Hence we consider selecting the lag length by maximizing the marginal likelihood, as we did before for the hyperparameter λ_1. At each forecast origin, we set

DOI: 10.1002/jae



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	1.4062	1.5519	1.2604
Median	0.8027	0.9665	0.6019
Min	-0.4993	-0.4993	-0.1603
Max	8.1932	6.5328	8.1932
%>0	94.9074	98.1481	91.6667
%[-1%, +1%]	58.7963	50.9259	66.6667
%p<=0.05	28.7037	45.3704	12.0370

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	0.4946	0.5699	0.4193
Median	0.5104	0.5938	0.4594
Min	-3.4078	-3.4078	-2.8322
Max	2.3883	2.3883	2.0200
%>0	82.4074	82.4074	82.4074
%[-1%, +1%]	68.5185	65.7407	71.2963
%p<=0.05	25.9259	32.4074	19.4444

Figure 1. BVAR with optimally selected shrinkage parameter versus baseline BVAR. The top panel presents the percentage gains in relative mean squared forecast error of model M versus model B, computed as follows: $\left(1 - \text{RMSFE}_{i,h}^{M}/\text{RMSFE}_{i,h}^{B}\right) \times 100$. The bottom panel presents the percentage gains in the score of model M versus model B, computed as follows: $\left(\text{SCORE}_{i,h}^{M} - \text{SCORE}_{i,h}^{B}\right) \times 100$. Results in the bar charts are displayed for selected forecast horizons: 1, 3, 6, and 12 steps ahead. The thinner lines above the bars denote rejection of the null of equal forecast accuracy at the 5% level. The tables contain descriptive statistics on the percentage gains, based on *all* forecast horizons. Descriptive statistics include average, median, maximum, minimum, percentage of cases in which the percentage gains are above 0, percentage of cases in which the percentage gains are between -1% and $+1\%$, and the percentage of cases in which the forecasts from the competing models are statistically different according to the Diebold–Mariano (1995) test with the Harvey *et al*. (1997) adjustment

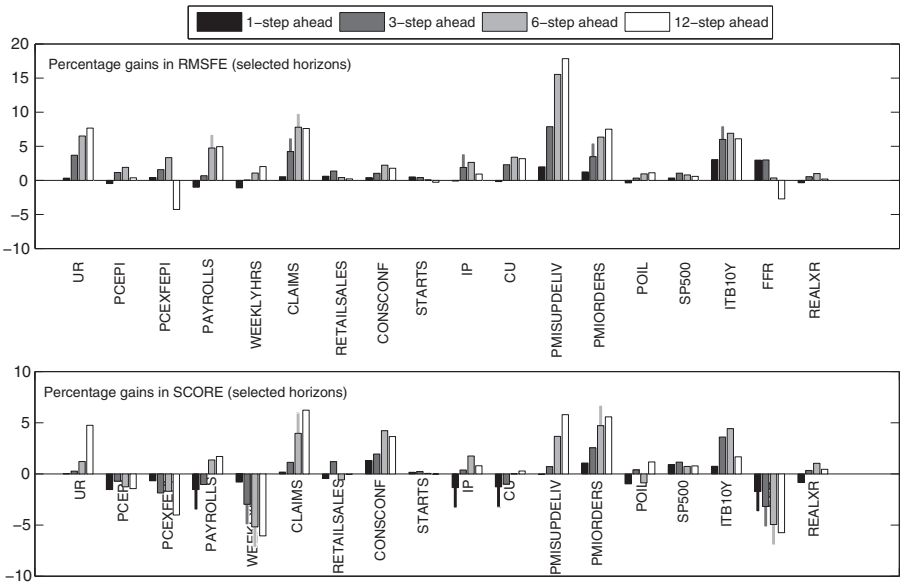
$$p_t^* = arg\max_{p} lnp(Y) \tag{23}$$

where we optimize over the grid $p = 1, 2, \ldots, 13$.[7]

For the US data the optimal lag length chosen with this method is equal to 3 in the first quarter of the estimation sample, 4 in the second quarter, and 13 in the second half of the sample. The results for the USA are reported in Figure 2, which provides the gains in RMSFE and SCORE of the specification with $p = p_t^*$ against the baseline specification with $p = 13$. It emerges that selecting the lag length

[7] We have also tried extending the maximum lag length up to 24 lags. Results for this case are virtually identical to those obtained with 12 lags and are available upon request.

DOI: 10.1002/jae



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	2.9626	2.2916	3.6336
Median	1.5771	1.2749	2.0241
Min	-4.2639	-1.0939	-4.2639
Max	19.7722	15.5368	19.7722
%>0	86.1111	89.8148	82.4074
%[-1%, +1%]	35.6481	37.9630	33.3333
%p <=0.05	17.1296	12.9630	21.2963

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	0.5297	0.1774	0.8820
Median	0.4081	0.2145	0.7909
Min	-6.4608	-5.1772	-6.4608
Max	6.4876	4.7162	6.4876
%>0	62.5000	58.3333	66.6667
%[-1%, +1%]	37.9630	43.5185	32.4074
%p <=0.05	25.9259	17.5926	34.2593

Figure 2. BVAR with optimally selected lags versus baseline BVAR. See notes to Figure 1

can be significantly beneficial for some variables, such as unemployment, the 10-year Treasury bill rate, the indices of supplier delivery times and new orders, and the claims for unemployment insurance. Losses are limited to a handful of variables, notably the FFR and hours in the case of density forecasts. Overall, optimizing over the lag length seems to help a little more than does optimizing over the shrinkage parameter. For example, across all variables and horizons, the median improvement in RMSFE associated with optimizing the lag is about 2.96 percentage points, while the median improvement in RMSFE associated with optimizing the overall shrinkage is about 1.4 percentage points.

We have also considered optimizing the rate at which longer lags are shrunk more strongly towards 0. To do so, we modified the standard deviation in (13) to $\frac{\lambda_1\lambda_2}{k^d}\sigma_i/\sigma_j$. The parameter d measures the rate of decay, and in the baseline specification is set to 1. The optimal rate of decay starts at 1.5 and then gradually declines to 1.2 for the later part of the sample. In terms of forecasting accuracy, optimizing over the lag decay d performs roughly the same as optimizing over the lag length p.

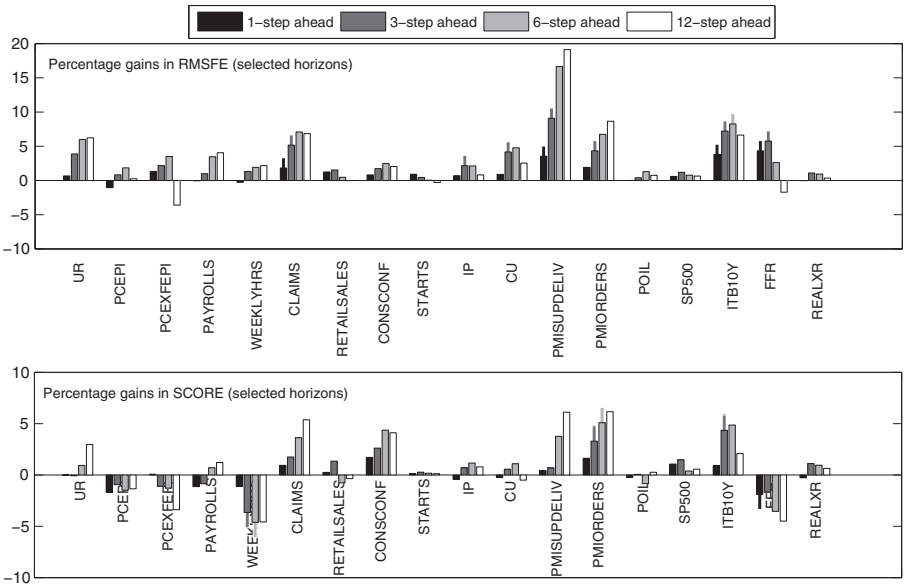
4.3. Simultaneous Choice of Hyperparameters and Lag Length

We now assess whether maximizing the marginal likelihood with respect to both the tightness parameter λ_1 and the lag length p can produce some additional gains. In order to do so, we compute the marginal likelihood using a limited range of values for λ_1 and p between 1 and 13. To streamline the computations, we consider a smaller grid of values for λ_1 than we did in the case in which we just optimized over λ_1, using a range of $\{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5\}$. Results in terms of lag selection

DOI: 10.1002/jae

do not change: the optimal lag length chosen with this method is equal to 3 in the first quarter of the sample, 4 in the second quarter, and 13 in the second half of the sample. However, once we allow for the lag length to change, we see some movements in the optimal tightness as well. In particular, the optimal tightness is equal to 0.2 in (roughly) the first half of the sample, and then becomes 0.15 in the second half. This likely happens because in the first part of the sample 3 or 4 lags are selected, so there are fewer regressors in the model and therefore less shrinkage is required. However, clearly the optimal parameter λ_1 still does not move much; therefore the forecasting results are very similar to those obtained by optimizing over the lag length, as shown in Figure 3. For example, the average RMSFE gains only increases to 3.3% from 2.96%.

To sum up, optimization in general is recommended. Maximizing the marginal likelihood using (22) requires little computational effort and has the important advantage that it will yield the optimal tightness for any dataset, regardless of its composition and size (in both the cross-sectional and time series dimension). The results presented here cannot guarantee that the tried-and-true values will work well in all cases, such as when the model includes 50 or 100 variables. That said, for models in common macroeconomic datasets such as those examined in this paper, the payoffs to optimization are often modest. We shall see in Section 10 that similar results are obtained for Canadian, French and UK data. Therefore, when optimization is viewed as too costly in terms of coding or computational time, the costs to using tried-and-true values in common data are fairly small. One reason is probably that the tried-and-true values have been established largely on the basis of forecasting performance.



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	3.3057	2.8990	3.7123
Median	2.0840	1.8705	2.1820
Min	-3.6090	-1.0439	-3.6090
Max	20.7923	16.6527	20.7923
%>0	88.8889	93.5185	84.2593
%[-1%,+1%]	31.0185	26.8519	35.1852
%p<=0.05	17.1296	26.8519	7.4074

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	0.7311	0.4940	0.9682
Median	0.5878	0.4170	0.7462
Min	-5.3271	-4.7006	-5.3271
Max	6.3255	5.0902	6.3255
%>0	67.1296	65.7407	68.5185
%[-1%,+1%]	38.4259	42.5926	34.2593
%p<=0.05	15.2778	12.9630	17.5926

Figure 3. BVAR with optimally selected tightness and lags versus baseline BVAR. See notes to Figure 1

DOI: 10.1002/jae

5. LEVEL VERSUS GROWTH RATES AND THE ROLE OF COINTEGRATION/UNIT ROOTS PRIORS

It is in principle unclear whether transforming variables into their growth rates can enhance the forecasting performance of the BVAR. Some researchers and practitioners prefer to leave variables in log levels and impose prior means of unit roots with additional priors on sums of coefficients (see, for example, Banbura *et al*., 2010; Giannone *et al*. 2012). One reason is that such a specification can better take into consideration the existence of long-run (cointegrating) relationships across the variables, which are omitted in a VAR in differences. On the other hand, Clements and Hendry (1996) show that in a classical framework differencing can improve the forecasting performance in the presence of instability. Diebold and Kilian (2000) show that, for variables with unit roots, forecasting accuracy can be improved by differencing. Hence this is another issue to be considered from an empirical perspective. As far as we know, there has been little published effort in the BVAR forecasting literature to compare specifications in levels versus differences. Following the Litterman (1986) tradition, some BVAR forecasting work uses models with variables in levels or log levels (e.g. Banbura *et al*., 2010; Giannone *et al*., 2010, 2012), while other work uses models in differences or growth rates (e.g. Del Negro and Schorfheide, 2004; Clark and McCracken, 2008; Koop, 2013).

Accordingly, we revisit the levels versus growth rates question. We estimate a version of the baseline BVAR in which many variables enter the model in growth rates (see Table I). As the transformed variables are likely to be stationary, we change the prior beliefs accordingly. In particular, the prior mean Φ^* in (13) is set to 0 for all variables, while we remove the unit root/cointegration priors.[8] We use 12 lags and set the overall shrinkage parameter λ_1 at 0.2.

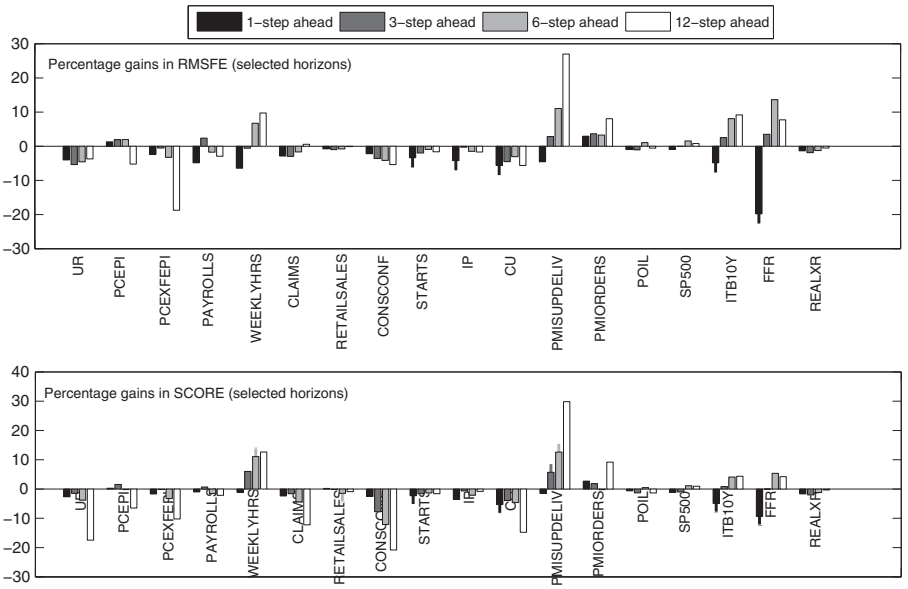
Results for the growth rate specification are displayed in Figure 4. On average over all variables and forecast horizons, the differences in the loss functions are small, with an average gain in RMSE of just 0.37% and an average loss in score of just 0.75%. When one looks at individual variables an interesting pattern emerges. Most of the variables feature a small increase in RMSFE when the forecasts are produced with the model in growth rates. Indeed, the model in growth rates outperforms the model in levels in only 36.6% of the combinations of variables and horizons. However, for a few variables at longer horizons (WEEKLYHRS, PMISUPDELIV, FFR and ITB10Y), the model in growth rates provides sizable forecasting gains. On average, these effects cancel out and overall we have that mean and median relative gains are very close to 0. Similarly, for average log scores, for most variables and horizons (67.6% of the cases), the model in levels performs better than the model in growth rates, although the median difference in scores is small. Again, though, for a few variables at longer horizons (e.g. WEEKLYHRS and PMISUPDELIV), the growth rates specification yields sizable improvements in scores.

For this growth rates specification we have also considered optimizing over the tightness hyperparameter and lag length. The optimal tightness changes only once, moving from a value of 0.25 in the first 40 estimation samples to 0.2 for all the remaining samples. The optimal lag length is instead fixed at 12 over all the samples. For this reason, in the growth rate case the results obtained by optimizing the hyperparameters are inevitably very close to those obtained with the baseline specification.

The role played by the inclusion of the sum of coefficients and initial dummy observation priors in the baseline levels specification also deserves investigation. If one decides to estimate the model in levels, then

[8] In our growth rates specification, all variables are assumed a priori to be stationary, and accordingly we follow studies such as Wright (2011) in setting all prior means to 0. However, while stationary, some variables can still be very persistent, and some studies (e.g. Clark, 2011) set the prior means of persistent variables to something like 0.8. Accordingly, we have run comparisons for the 18 and 7 variable models for the USA and the models for the other countries to check forecast accuracy for models in growth rates with different prior means for the AR(1) coefficients. Specifically, for the persistent variables in each growth rates model, we set the prior mean of each AR(1) coefficient to 0.8 instead of the 0 we use in the paper's results for the growth rates specification. At least in these checks the modified prior means yielded results quite similar to those reported in the paper. On average across variables and forecast horizons, the accuracy of the growth rates model with a mixture of AR(1) prior means of 0 and prior means of 0.8 was essentially the same as the accuracy of the model with prior means of 0 for all coefficients. These results apply to both recursive and rolling estimation schemes.

DOI: 10.1002/jae



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	0.3697	-0.4705	1.2099
Median	-0.7938	-1.0268	-0.5534
Min	-19.7989	-19.7989	-18.7466
Max	26.9899	13.6283	26.9899
%>0	36.5741	32.4074	40.7407
%[-1%,+1%]	23.1481	19.4444	26.8519
%p<=0.05	3.2407	5.5556	0.9259

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	-0.7537	-0.5762	-0.9313
Median	-1.2070	-1.1750	-1.2796
Min	-20.8116	-12.1092	-20.8116
Max	29.8323	12.6247	29.8323
%>0	32.4074	30.5556	34.2593
%[-1%,+1%]	21.2963	25.9259	16.6667
%p<=0.05	15.2778	12.0370	18.5185

Figure 4. BVAR in growth rates versus baseline BVAR (in levels). See notes to Figure 1

these priors implement the belief that there is unit-root/cointegration in the system, which is a typical feature of macroeconomic datasets such as the one considered here. Sims (1993) shows examples that 'without such elements in the prior, fitted multivariate time series models tend to imply that an unreasonably large share of the sample period variation in the data is accounted for by deterministic components'. In order to check for this effect we have also re-estimated our BVAR in levels omitting the sum of coefficients and initial dummy observation prior. We do not report the results here, for brevity, but they are available in the supporting information, Figure 11. The main finding is that once one does that, the model in growth rates systematically dominates the model in levels. This is evident in RMSFEs (relative to the baseline model) that are much higher for the model in levels without these priors (by an average of about 11%) than for the model in growth rates. This suggests that, while the model in growth rates does not need by construction the cointegration/unit root prior (because it is stationary), when the model is estimated in levels such priors do help and should be included, consistent with evidence in Banbura *et al*. (2010).

Finally, it is worth mentioning that, even when working with larger VARs, forecasters might be interested in forecasting only a few variables in the system, e.g. inflation and output growth. In this paper, selections of lag length and prior hyperparameters were based on all of the variables. While we do not pursue this here to save space, a promising direction for further research is to optimize only the marginal likelihood of the variables of interest.[9]

[9] However, Jarocinski and Mackowiak (2011) argue for focusing on model fit measures for full sets of variables.

DOI: 10.1002/jae

To sum up, we find that specifications in levels and growth rates produce on average comparable forecasts. For most variables, the model in levels produces more often (i.e. for more variables) gains in point and density forecasting, but the model in growth rates performs particularly well for selected variables. The good performance of the model in levels deteriorates substantially if one removes the unit roots/cointegration priors. Therefore, if one wants to use the traditional Minnesota prior specified in (13), then we recommend working in differences. For a VAR in levels, the traditional Minnesota prior should be combined with the dummy observation/sum of coefficients priors in (14) and (15). Overall, the latter specification should be preferred.[10]

6. ALTERNATIVE MULTI-STEP FORECASTING APPROACHES

In this section we compare the simulation-based approach with multi-step forecasting used so far versus some alternative methods that can considerably reduce the computational burden.

6.1. Full Simulation versus Approximation (Iterated vs. Pseudo-iterated Approach)

As discussed, for the standard N-IW prior, closed-form solutions are available for the marginal posterior of the VAR coefficients. These would naturally provide closed-form solutions for the one-step-ahead forecasts. However, for multi-step forecasting (and also for impulse response analysis) the fact that coefficients enter nonlinearly implies that simulation methods are needed. The posterior distribution of the h-step ahead forecast is a nonlinear function of Φ and therefore can only be obtained by simulation. For example, using the companion form notation given in (9), the posterior mean of the forecasts would be given by

$$\hat{y}_{t+h} = \frac{1}{m}\Sigma_{l=1}^{m}\left[\mathbf{s}\cdot\left(\Phi_l^+\right)^h x_{t+1}\right] \tag{24}$$

where $\left(\Phi_l^+\right)^h x_{t+1}, l = 1, \ldots, m$, is a collection of m simulated forecasts based on m draws from the marginal of Φ, which are obtained by using (12) and the computational efficiency of (19). We label this approach, which is used in our baseline specification, the 'iterated' approach.

Therefore, the production of multi-step point forecasts using simulation and then equation (24) will require m times more computations (each of order $M^3 + N^3$) than producing the one-step-ahead forecasts via the closed-form solution in (17). This is not a problem if one considers a single estimation of a BVAR, but of course if one is back-testing several different models by evaluating historical forecasts it can increase CPU time requirements substantially. As an alternative, one can choose to approximate the results by just integrating out the uncertainty in the coefficients and then using the posterior mean of the coefficients to produce posterior means of the multi-step forecasts. In this case the multi-step point forecast is computed as

$$\hat{y}_{t+h} = \mathbf{s}\cdot\left(\bar{\Phi}^+\right)^h x_{t+1.} \tag{25}$$

This method has been used, for example, by Banbura *et al.* (2010), and we label it the 'pseudo-iterated' approach. Of course this approach has a clear computational benefit but it is, strictly speaking, inaccurate as it ignores the nonlinearity inherent in multi-step forecasting. In this section we assess the effects of

[10] Studies such as Clark (2011), Osterholm (2008) and Wright (2011) have shown that the performance of models with variables transformed for stationarity can be improved significantly with the use of the steady-state prior developed in Villani (2009). As Villani's estimation approach involves Gibbs sampling, it is significantly more computationally demanding than a model specified to permit a Normal-inverted Wishart prior and posterior. Accordingly, in light of our focus on monthly data and 18 variables, we do not pursue the steady-state prior in this paper.

DOI: 10.1002/jae

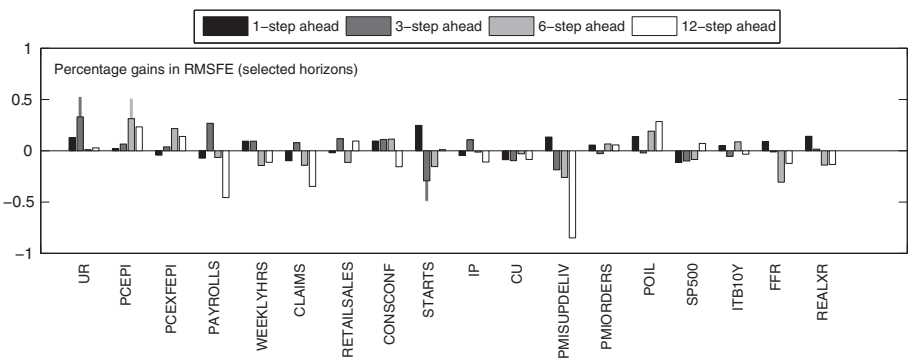
ignoring these nonlinearities on the precision of point forecasts, and we find that the cost is very small (it is not likely to be small for density forecasts).

In some cases there can be little choice but to use the 'pseudo-iterated' approach. If one departs from the N-IW conjugate prior, which ensures both a closed-form solution for the joint posterior density of the parameters and a particularly convenient Kronecker structure of the posterior coefficient variance matrix, the computational costs involved in simulating the joint posterior distribution of the parameters increase rapidly with the dimension of the system, because to draw a sequence of Φ one must resort to manipulation of an $MN \times MN$ variance matrix without the efficient Kronecker structure of (12) and (19). The computational costs rise sharply with the number of lags and, in particular, number of variables. Other studies such as Kadiyala and Karlsson (1997) and Karlsson (2012) have stressed the computational challenges associated with specifications that depart from the Kronecker structure.

Moreover, the use of the 'pseudo-iterated' approach may be obliged for some practitioners relying on common software packages such as RATS and Eviews that do not provide simple, direct commands for simulating BVARs. In these packages, commands produce posterior moments, but do not permit direct simulation of the posterior distributions. Instead, for simulation, users must be able to write their own programs, as would also be the case for packages such as Matlab.

Our take on these considerations is that they certainly warrant looking at the accuracy of forecasts obtained by methods that do not involve simulation. So we turn now to comparing the results obtained by using the pseudo-iterated approach with those resulting from the proper simulation-based iterated approach, using in both cases the benchmark specification with 13 lags and fixed tightness.

Results of this experiment are contained in Figure 5 and clearly indicate that the gains from the simulation-based approach are negligible. Therefore, if one is only interested in point forecasts, the loss from the quick and easy 'pseudo iterated' approach is small. On the other hand, if the focus is on multi-step density forecasts, the pseudo-iterated approach would not help from a computational point of view. Indeed, while one could compute a set of desired quantiles from an (approximate) predictive density based on an iteration of equation (17), the proper production of a whole density forecast would still require simulation methods.



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	-0.0192	0.0082	-0.0466
Median	-0.0021	0.0070	-0.0064
Min	-0.8495	-0.4758	-0.8495
Max	0.3443	0.3305	0.3443
%>0	49.5370	50.9259	48.1481
%[-1%, +1%]	100	100	100
%p<=0.05	7.8704	6.4815	9.2593

Figure 5. BVAR estimated with pseudo-iterated approach versus baseline BVAR. See notes to Figure 1

DOI: 10.1002/jae

6.2. Direct Forecasting Approach

Another way to overcome the problem of nonlinearity in the multi-step forecasts is to use the so-called direct approach. Consider the following VAR:

$$y_t = \Phi_{c,h} + \Phi_{1,h} y_{t-(h-1)-1} + \Phi_{2,h} y_{t-(h-1)-2} + \ldots + \Phi_{1,h} y_{t-(h-1)-p} + \varepsilon_t. \quad (26)$$

Note that in the above model the vector y_t is regressed directly onto y_{t-h} and *p-1* additional lags, and that for each forecast horizon h a different model is employed. Such an approach is known as 'direct' forecasting, and it focuses on minimizing the relevant loss function for each forecast horizon, i.e. the h-step-ahead forecast error. Such an approach has been implemented in a Bayesian framework by Koop (2013), for example. The approach of our baseline specification, namely regress y_t onto y_{t-1}, $\ldots, y_{t-p}$ and then compute recursively the h-step-ahead forecasts, is known as 'iterated' forecasting or 'powering up'. For a discussion and a comparison of these alternative methods in a classical context see, for example, Marcellino *et al*. (2006) and Pesaran *et al*. (2011).

In brief, generally, the powering-up approach is more efficient in a classical context, as the used estimators are equivalent to maximum likelihood, under correct model specification. But it is dangerous in the presence of misspecification, because in general the misspecification will inflate with the forecast horizon when the forecasts are computed recursively. In addition, the direct approach implies that the h-step-ahead forecast is still a linear function of the coefficients (because a different model is used for each forecast horizon), while in the traditional powering-up approach the multi-step forecasts are highly nonlinear functions of the estimated coefficients. As a result, there is an exact closed-form solution for the distribution of the h-step-ahead forecasts computed using (26), while computing the forecasts resulting from the powering-up strategy requires the use of simulation methods, as discussed above.[11]

Since the final forecasts are forecasts of the growth rates of the variables, getting density forecasts for growth rates starting from forecasts based on the levels specification is not straightforward. Accordingly, in evaluating forecasts based on a direct approach to estimation and forecasting, we use as a benchmark forecasts from a model in what we have referred to as growth rate form, rather than levels form; we obtain the benchmark forecasts by simulation and iteration.

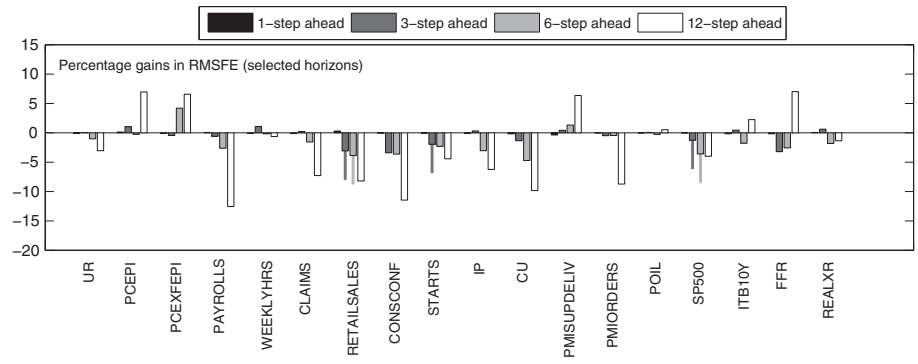
In the direct case, the optimal lag length (based on the marginal likelihood) is 12 over the whole sample. In this case, as a different model is estimated at each forecast horizon, a separate optimal shrinkage parameter is estimated for each horizon. Also in this case the shrinkage parameter does not vary a lot. At one step ahead, the optimal shrinkage parameter is 0.2 in 85% of the samples and 0.25 in the remaining ones. For the two-step-ahead horizon, it is 0.25 in 77% of the samples and 0.15 in the remaining ones. For horizons of three months or more, the optimal value is 0.2 almost in each sample, while the value 0.25 is chosen in a few other cases (generally around 1% of the cases and never more than 5% of the total samples).

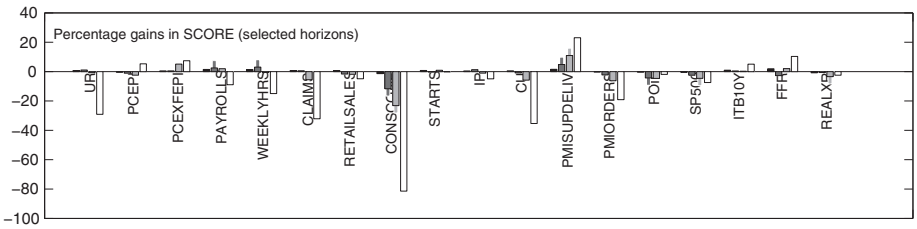
With these results, the use of optimal tightness and lag length offer little benefits with respect to the baseline model (in growth rates), and therefore we concentrate on the fixed shrinkage and lag case. Results for the other cases are very similar and are available upon request. The outcome of the comparison of the direct forecasting method against the baseline growth rates specification is reported in Figure 6.

In this case there is no difference by construction for one-step-ahead forecasts, except for very small differences that arise due to the simulation of the baseline model in growth rates. For horizons shorter than six steps ahead, there is little loss in using the direct approach, with an average loss below 1% both

[11] Admittedly, however, the closed-form solution obtained with a direct forecasting approach assumes the error terms of the model are serially uncorrelated, which will not actually be the case with forecast horizons of more than one period. We follow other studies such as Jacobson and Karlsson (2004), Koop (2013) and Wright (2009) in applying direct methods to multi-step forecast horizons.

DOI: 10.1002/jae





Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	-1.5858	-0.8614	-2.3102
Median	-1.4034	-0.2993	-3.0039
Min	-12.5378	-5.5387	-12.5378
Max	7.0107	4.2022	7.0107
%>0	28.7037	32.4074	25
%[-1%, +1%]	30.5556	52.7778	8.3333
%p<=0.05	19.9074	10.1852	29.6296

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	-3.6456	-0.8788	-6.4123
Median	-1.1933	-0.2972	-3.4425
Min	-81.4222	-23.1846	-81.4222
Max	23.0603	10.9465	23.0603
%>0	36.1111	47.2222	25
%[-1%, +1%]	24.0741	38.8889	9.2593
%p<=0.05	15.7407	17.5926	13.8889

Figure 6. BVAR estimated with direct approach versus baseline BVAR (models in growth rates). See notes to Figure 1

for the RMSFE and SCORE loss functions. As the forecast horizon increases, the direct method is outperformed by the baseline approach, especially for the SCORE loss function, with an average loss of 6.41%. For the RMSFE loss function, the average loss is only 2.31%. There are variables for which the choice between the direct and iterated approaches makes a sizable difference at long horizons, such as payrolls, claims, retail sales, consumer confidence and housing starts.

Overall, the pattern emerging from the findings documented in this section is that simulations are not strictly required to get good multi-step point forecasts from Bayesian VARs. The pseudo-iterated method works as well as the fully iterated approach, with virtually no losses in terms of point forecasts, and it is much easier and faster, in particular in a recursive context. For models with variables transformed for stationarity, in point forecast accuracy, the direct method performs comparably to the iterated approach. For density forecasting, the direct method seems suboptimal with respect to the fully iterated approach—more so (as one would expect) at longer forecast horizons than shorter horizons.

7. LITTERMAN PRIOR AND CROSS-VARIABLE SHRINKAGE

The baseline specification we have considered so far is a Normal-inverted Wishart conjugate prior which features the same sample mean for the VAR coefficients of the prior proposed by Litterman (1986) and which is known as Minnesota prior. However, the N-IW prior of our baseline specification differs in three respects. First, in Litterman's original implementation the unit root/cointegration priors

DOI: 10.1002/jae

were not considered (these were introduced by Doan *et al*., 1984, and Sims, 1993). We have analyzed the role played by such priors in Section 5 and concluded that it is indeed relevant in models in levels.

Second, in the original Litterman implementation, the hyperparameter λ_2 in the prior variance is set to a value smaller than 1, which puts additional shrinkage on the lags of all the variables other than the dependent variable of the *i*th VAR equation. Litterman (1986) sets this parameter to a value smaller than 1 in order to capture the idea that, at least in principle, these lags should be less relevant than the lag of the dependent variable itself. This modification implies that the Kronecker form for the coefficient variance matrix breaks down and as a consequence one can only derive the conditional posteriors, while to draw from the joint posterior of the coefficients and error variance matrix one needs to implement MCMC (Gibbs) sampling. Gibbs sampling has poorer mixing properties than the simple MC integration required for the N-IW case as MCMC methods produce autocorrelated draws. Moreover, a Gibbs sampling algorithm would require at each iteration the manipulation of $MN \times MN$ matrices to derive the conditional posterior mean of the coefficients and to perform a random draw from the conditional posterior.

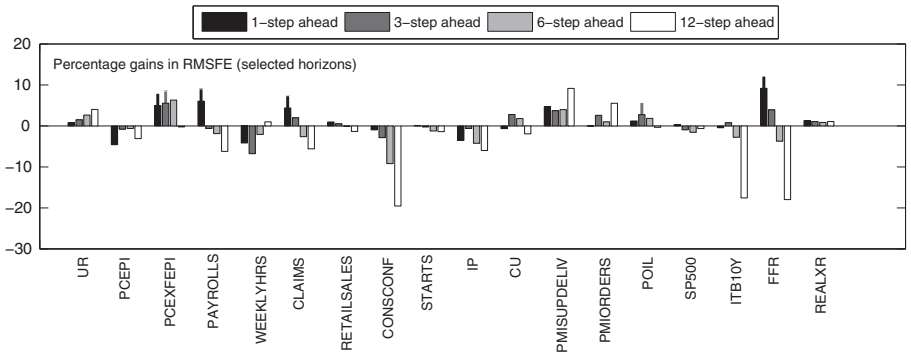
The third difference in the Litterman (1986) approach arises precisely because of the difficulty of estimating a large system when the cross-variable shrinkage is imposed. To overcome this, Litterman (1986) treats the error covariance matrix as fixed and diagonal and estimates it in a preliminary step. This assumption means that the model can be estimated with ridge regression on an equation-by-equation basis. In contrast, in our baseline N-IW prior, the covariance matrix is sampled from an inverted Wishart, calibrated so that its expected value coincides with the fixed diagonal matrix of Litterman (1986).

While the pioneering work of Litterman (1986) suggested it was useful to have cross-variable shrinkage, it has become more common to estimate larger models without cross-variable shrinkage, in order to have a Kronecker structure that speeds up computations and facilitates simulation. Still, pre-programmed Bayesian capabilities in programs like RATS include an option for cross-variable shrinkage.

To assess these specification choices, Figure 7 provides results for our Litterman specification of a model in levels, using cross-variable shrinkage of $\lambda_2 = 0.2$. To be able to compare density results, we simulated forecasts for the Litterman specification using the posterior normal distribution for each equation's coefficients, treating each equation independently per Kadiyala and Karlsson (1997) and holding the error variance matrix fixed (and diagonal). The Litterman approach that uses both cross-variable shrinkage and a diagonal error variance matrix fares on average slightly worse than the baseline model. The average losses are rather small: about 0.92% for the RMSFE and 1.41% for the SCORE loss function. These apparently clear-cut results hide an interesting feature, which becomes apparent when one looks at the role played by the cross-variable shrinkage versus the diagonal error variance matrix. To shed light on this, we estimated the BVAR using the Litterman estimation approach but setting $\lambda_2 = 1$. This case is in between the Litterman approach and the baseline model: it can be thought of as the Littermann approach with no cross-variable shrinkage, or as the baseline model with a diagonal variance matrix. Results for this case (available in the supporting information, Figure 12) show a clear deterioration with respect to the baseline model, with an average loss of 10.57% in RMSFE and 24.87% in SCORE. Moreover, the model is outperformed by the baseline specification in 100% of the cases for the RMSFE and 97% of the cases for the SCORE. Therefore, it seems that, by itself, the use of a diagonal variance matrix in the baseline specification reduces forecast accuracy, while the use of cross-variable shrinkage in the Litterman approach improves accuracy. On balance, these two effects offset each other.[12]

[12] We have also tried a modification of the Litterman approach that partially deals with the problem of the diagonal error variance matrix by allowing it to be non-diagonal when the error term is drawn (but still being diagonal when the coefficients are drawn). This indeed improves the average log scores.

DOI: 10.1002/jae



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	-0.9191	0.4612	-2.2995
Median	-0.3943	0.4831	-0.9365
Min	-19.5470	-9.1974	-19.5470
Max	9.1722	9.1722	9.1500
%>0	44.4444	54.6296	34.2593
%[-1%, +1%]	29.6296	30.5556	28.7037
%p <=0.05	6.9444	9.2593	4.6296

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	-1.4083	-1.0216	-1.7951
Median	-0.9873	-1.0030	-0.0895
Min	-25.8105	-13.8342	-25.8105
Max	9.1846	7.0576	9.1846
%>0	44.4444	40.7407	48.1481
%[-1%, +1%]	14.3519	15.7407	12.9630
%p <=0.05	23.6111	23.1481	24.0741

Figure 7. BVAR with Litterman prior versus baseline. See notes to Figure 1

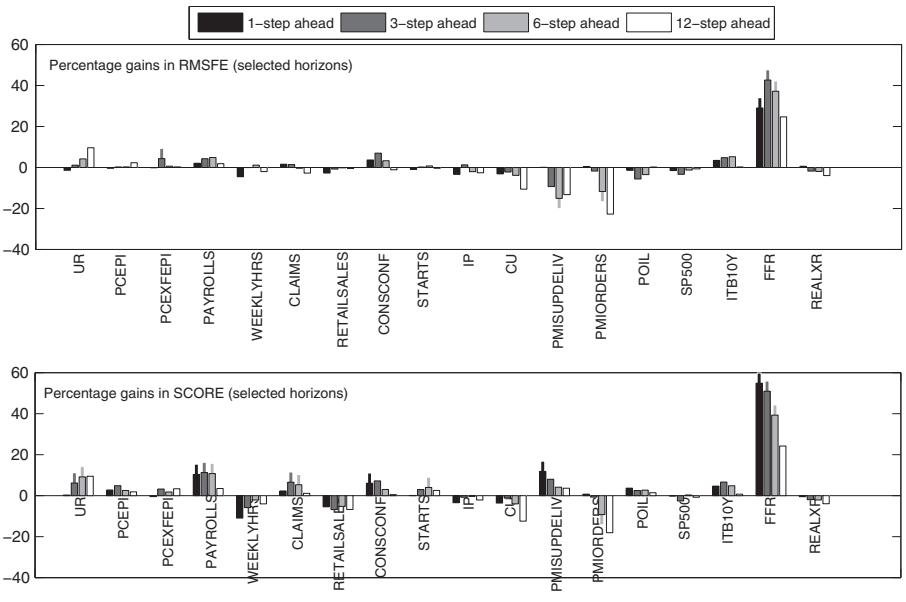
To sum up, the imposition of a diagonal error variance matrix is detrimental to forecasting accuracy, especially in density forecasting. Imposing cross-variable shrinkage provides a benefit, but on average such benefit is offset by the cost of imposing the diagonality in the variance of the system. Finally, while imposing cross-variable shrinkage and allowing a non-diagonal variance matrix is possible in principle, estimation should be performed via Gibbs sampling and quickly becomes difficult from a computational perspective (see, for example, Karlsson, 2012).

8. ROLLING VERSUS RECURSIVE ESTIMATION

There is a long debate in the forecasting literature on the relative merits of rolling versus recursive estimation. The former can be more robust in the presence of structural breaks, while the latter can be more efficient. Hence we now assess their performance in our context and evaluate whether the other results we have obtained so far are robust to the choice of the estimation method.

To start, in Figure 8 we compare point and density forecasts from recursive and rolling estimates of the benchmark specification, taking the recursive case as the benchmark. The rolling estimates use a window of 11 years of data, corresponding to the size of the sample used to generate the first forecast observation in the recursive scheme. On average, the two methods perform broadly similarly, particularly in terms of RMSFEs. But looking at the percentage of cases in which a given method outperforms the other, it appears that the rolling method performs relatively better in density forecasting, while the recursive method performs relatively better in point forecasting. We interpret this finding as

DOI: 10.1002/jae



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	0.6604	1.6721	-0.3513
Median	-0.2057	0.1102	-0.5028
Min	-22.7504	-15.0669	-22.7504
Max	42.6381	42.6381	34.2417
%>0	47.2222	50.9259	43.5185
%[-1%,+1%]	28.7037	26.8519	30.5556
%p<=0.05	13.8889	12.9630	14.8148

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	2.8796	4.3679	1.3914
Median	1.8493	2.4997	1.4470
Min	-18.0863	-10.9365	-18.0863
Max	54.8665	54.8665	35.3132
%>0	61.5741	62.0370	61.1111
%[-1%,+1%]	12.5000	13.8889	11.1111
%p<=0.05	24.5370	24.0741	25

Figure 8. BVAR estimated with rolling window versus baseline (recursive window). See notes to Figure 1

possibly related to time-varying volatility in the errors. Indeed, when one considers density forecasts, the assumption of homoskedasticity and Gaussianity is restrictive and drifting volatilities likely matter. Introducing stochastic volatility in a VAR of this dimension creates substantial computational problems and goes beyond the scope of this paper, but it is considered in Carriero *et al.* (2011).[13] It is also worth mentioning that rolling estimation improves substantially the FFR forecasts under both loss functions, a finding that might be explained by the ability of rolling forecasts to 'forget' about previous policy regimes.

We now assess the robustness of the previous findings related to the usefulness of optimal selection of shrinkage and lag length and cross-variable shrinkage. To save space, we present a summary of the results, with full details available upon request. When we optimize the overall shrinkage hyperparameter λ_1, the optimal setting moves a bit more compared to the recursive case, when it was steady at 0.15. In particular, the parameter still remains at 0.15 for the majority (86.4%) of the samples, but in the other cases it moves up to 0.2. The resulting forecasts, are however, fairly close to those based on a fixed tightness, confirming what we found before. When we optimize the lag order, the selected lag is much more variable than in the recursive case. The optimal lag starts at 3, rises to 6 or 7 in the mid 1990s, bounces around to levels as high as 12, and then ends the sample at 9. Consistent with the recursive results, for models estimated with a rolling sample, optimizing the lag improves accuracy (compared to the case of a fixed lag). For example, RMSFE is lower with the optimal lag

[13] Koop and Korobilis (2012) allow a computationally simpler form of time-varying volatility in large BVARs, through an exponentially weighted moving average filter.

DOI: 10.1002/jae

in about 82% of the horizon and variable combinations, by an average of 2%. Finally, the results on cross-variable shrinkage are in line with what we documented in the case of recursive estimation.

In summary, rolling estimation can slightly improve the density forecasts, but not the point ones, with respect to recursive estimation. However, the other main findings of the paper in terms of small but positive gains coming from hyperparameter and lag length selection are overall confirmed also with rolling estimation.

9. VAR SIZE

While a number of studies have found forecast accuracy improves with larger datasets, it is not necessarily the case that more is always better. For example, Boivin and Ng (2006) suggest that pre-selecting the variables that are included in a factor model according to their relationship with the target variable of interest can improve the forecasting precision. Similarly, Banbura *et al*. (2010) show that a medium-scale BVAR of about 20 variables delivers often more accurate forecasts than large BVARs. Koop (2013) shows that the forecasting performance increases with size, but only up to about 20 variables.

Therefore, we now assess the forecasting performance of a smaller-scale BVAR for a subset of the variables of interest, comparing it to that of our benchmark medium-sized VAR. Then, we consider whether the findings on the role of the BVAR specification choices remain valid for the smaller system.

We focus on the following seven variables: unemployment rate (UR), core PCE price index (PCEXFEPI), nonfarm payroll employment (PAYROLLS), nominal retail sales (RETAILSALES), single-family housing starts (STARTS), industrial production (IP) and the federal funds rate (FFR).

Results based on the baseline specification for both VARs are displayed in Figure 9. It is interesting to note that while on average the small system seems to produce slightly better forecasts with respect to the large system, for most variables and horizons (38.1% for the RMSFE and 29.8% for the SCORE), the small system is actually less accurate. The similar average performance of the two models is indeed driven by the particularly good performance of the small system in forecasting the FFR. If one does not consider this variable, then the large system produces better forecasts in most cases.

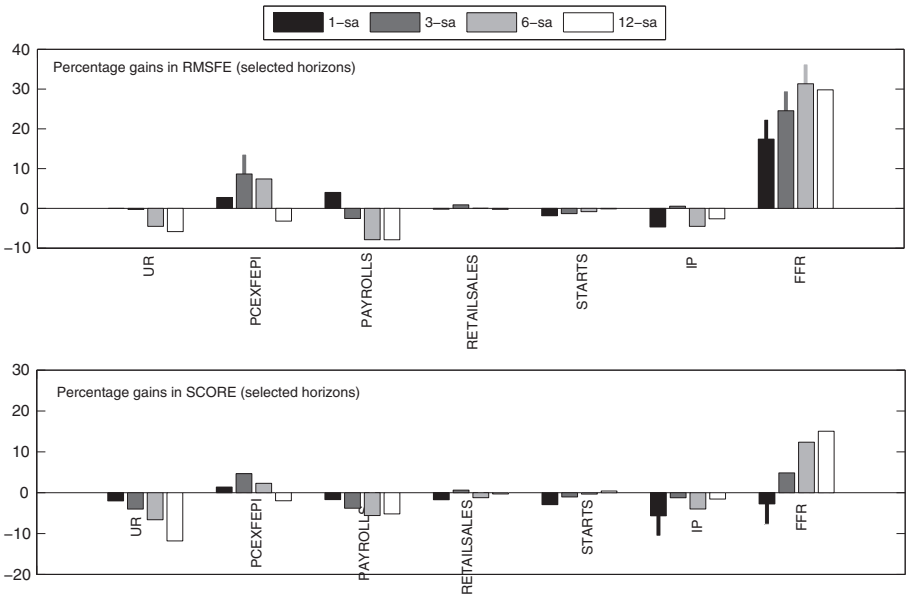
Let us now assess the robustness of the results obtained using the baseline VAR with 18 variables with respect to a set of specification choices, specifically, optimal tightness and lag length, levels versus growth rates, and iterated, pseudo-iterated and direct approach. To save space, we do not report the detailed results, but they are available upon request.

With the seven-variable VAR, the optimal lag length selected is 7 in the first quarter of the sample, 8 in the second quarter and 13 in the second half. It seems that, with some variables dropped out, the smaller model needs longer lags to soak up the associated dynamics. As the selected lag length is always quite high, the gains in choosing optimally the lag length are limited. The mean and median of the loss functions are very similar to those computed using the 13-lag specification, and the percentage of cases where the optimal selection pays off is close to 50%. When considering optimal tightness (λ_1) selection, the optimal tightness comes out at 0.2 for all forecast origins, up (implying less shrinkage) a bit from the 0.15 that proved optimal in the 18-variable model. The similarity of the shrinkage selections for the seven-variable and 18-variable models is consistent with the conventional wisdom that a setting of 0.2 generally works well. As in the 18-variable case, with just seven variables in the model the mean and median loss function values with optimal shrinkage are very similar to those obtained with fixed tightness. Moreover, the percentages of instances where a variable is better forecast by the model with optimal tightness is 43%, so in this case selecting optimally the tightness slightly reduces forecast accuracy on average. These results do not change much when tightness and lag length are jointly optimized.

With regard to the effect of levels versus growth rates and the use of iterated, pseudo-iterated and direct forecasting approaches, results are in line with those obtained with the baseline 18-variable specification.

With regard to the direct approach, the advantage of the 18-variable model over the seven-variable model is smaller. On average, the larger model is more accurate, but the difference between the two

DOI: 10.1002/jae



Percentage gains in RMSFE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	2.5896	3.3956	1.7837
Median	-0.3522	-0.0515	-0.8480
Min	-11.3216	-7.9728	-11.3216
Max	31.6865	31.3349	31.6865
%>0	38.0952	47.6190	28.5714
%[-1%, +1%]	33.3333	35.7143	30.9524
%p <=0.05	16.6667	19.0476	14.2857

Percentage gains in SCORE descriptive statistics (all horizons)

	All horizons	Hor.<=6	Hor.>6
Mean	-0.8078	-0.6656	-0.9500
Median	-1.6264	-1.2338	-1.8837
Min	-11.8156	-6.6313	-11.8156
Max	15.0374	12.3616	15.0374
%>0	29.7619	30.9524	28.5714
%[-1%, +1%]	21.4286	19.0476	23.8095
%p <=0.05	9.5238	7.1429	11.9048

Figure 9. BVAR based on seven variables versus baseline BVAR (based on 18 variables). See notes to Figure 1

models is smaller than under the pseudo-iterated and iterated approaches. In fact, at longer forecast horizons, for many variables the smaller model becomes slightly more accurate than the larger model. This might indicate that the seven-variable VAR is somewhat misspecified with respect to the 18-variable VAR, and the direct approach is better suited to handle such misspecification.

10. RESULTS FOR OTHER COUNTRIES

To make sure our conclusions have broad applicability, we extend our analysis to three more countries: Canada, France and the UK. For each country we have collected a dataset composed of nine variables since the entire set of variables used for the US analysis is not available for each country, or at least not for a sufficiently long time span. The variables, together with their transformations, are described in Table I (panel B).

The estimation and forecast samples are comparable to those for the USA. Specifically, for Canada and France we use data ranging from January 1971 to May 2010. The first estimation sample is February 1972 to December 1983, and then the estimation sample expands with the recursive scheme, ending in May 2009. The forecast period for these countries ranges from January 1984 to May 2010. For the UK the sample is slightly shorter, starting in January 1975 and ending in March 2010. The first estimation window is February 1976 to December 1987, the last is January 1975 to March 2009, and the forecast sample is January 1988 to March 2010.

The benchmark specification is as for the USA; therefore with a fixed lag length and tightness, variables in levels (with unit root and sums of coefficients priors), full iteration to obtain h-step-ahead

DOI: 10.1002/jae

forecasts and recursive estimation. We then assess the role of optimal selection of the lag length and prior tightness; growth versus level specification; alternative methods for computing h-step-ahead forecasts (pseudo-iterated and direct); cross-variable shrinkage; and rolling versus recursive estimation. In most cases we consider both point and density forecasts.

In the interest of space, here we briefly summarize the results, focusing especially on the comparison with the findings obtained for the USA. More detailed results can be found in the online Appendix (supporting information).

The results on the choice of the tightness parameter and lag are broadly in line with those for the USA, but with some small differences. For two of the three other countries (Canada and France), optimizing the shrinkage yields slightly larger and more consistent gains than does optimizing the lag, while lag shrinkage performs relatively better for the USA and UK. For example, in the case of Canada, optimizing shrinkage improves RMSFEs relative to the baseline in 89.8% of the variable and horizon combinations, with a median gain of 1.15%; optimizing the lag improves RMSFEs in 66.7% of cases, with a median improvement of 0.73%. For the UK, optimizing shrinkage also improves RMSFEs in 89.8% of cases, with a median gain of 0.83%; optimizing the lag improves RMSFEs in 87.0% of cases, with a median improvement of 1.74%. Optimizing both shrinkage and lag together does not add any additional RMSFE gains. Therefore, the overall message remains that optimizing over the tightness hyperparameter and lag length can be helpful for the majority of variables and forecast horizons, though for most variables and horizons the gains are limited.

The comparison of levels versus growth rates mostly—although not entirely—resembles that for the USA. For Canada and France, like the USA, forecast accuracy is similar for models in levels and growth rates, when measured either by RMSFE or predictive score. However, for the UK the levels specification has a clearer advantage over the growth specification, with average RMSFE and score gains of about 2% and 5%, respectively, and a better performance in about 60% of cases. Overall, with respect to the USA, there seems to be less support for specifications in growth rates.

Turning to the findings on the iterated versus pseudo-iterated approach, for the US we found clear evidence that the two methods produce virtually the same results in terms of point forecasts, supporting the use of the much quicker pseudo-iterated approach. These results are resoundingly confirmed. For Canada, France and the UK, the RMSFEs never differ by 1% or more.

As for the comparison between the direct and iterated approach (for models in what we refer to as growth rate form), the results for other countries are broadly similar to those for the USA: there is little to be gained by the use of direct multi-step estimation and forecasting. For Canada, France and the UK, RMSFEs are on average slightly higher with the direct approach than the baseline specification (in levels, with iterated forecasts), with RMSFEs that are higher in most variable-horizon combinations. The direct approach also yields slightly lower scores, although the percentage of cases in which it is less accurate than the baseline is smaller in terms of scores than in terms of RMSFEs.

With regard to the comparison between the Litterman specification prior and the baseline specification we confirm that the cross-variable shrinkage, when coupled with a fixed and diagonal error variance matrix, does not pay a lot, since the fraction of cases in which the simpler baseline specification forecasts better is often above 50%, and it yields average gains in the range of 1–4%.

Finally, about rolling rather than recursive estimation, the former performs slightly better than for the USA, in particular for density forecasting. In terms of RMSFE the largest average gains are about 3% for the UK, with a value of 2% for Canada and France. The gains in terms of predictive score are instead in the range 5–11%, with 71–91% of cases where the rolling score is better than the recursive one. As mentioned in the case of the USA, rolling estimation could provide more robustness in the presence of parameter time variation, and this seems to matter more for density forecasts.

DOI: 10.1002/jae

11. CONCLUSIONS

In this paper we discuss how a set of specification choices affects the forecasting performance of Bayesian VARs. We adopt as a benchmark a common specification in the literature, a Bayesian VAR with variables entering in levels and a prior modeled along the lines of Sims and Zha (1998). We then consider optimal choice of tightness, of lag length and of both; consider the relative merits of modeling in levels or growth rates; compare alternative approaches to multi-step forecasting (direct, iterated and pseudo-iterated); and discuss the treatment of the error variance and of cross-variable shrinkage.

To ensure our results have broad applicability, we check their robustness to the choice of the sample size (the time series dimension of the VAR) by comparing rolling with recursive estimation; to the VAR size (the cross-sectional dimension of the VAR) by analyzing a subset of seven of the 18 US variables; and to the VAR composition, by repeating the analysis for some other datasets—specifically, data for Canada, France and the UK.

We obtain a large set of empirical results. We can summarize them by saying that we find very small losses (and sometimes even gains) from the adoption of BVAR modeling choices that make forecast computation quick and easy, in particular for point forecasts. An approach that works well is to specify a Normal-inverted Wishart prior along the lines of Sims and Zha (1998) on the VAR in levels, preferably optimizing its tightness and lag length. Optimizing over the lag length tends to be more helpful (i.e. providing relatively larger gains) than optimizing the tightness. For the accuracy of point forecasts, there proves to be essentially no payoff to using simulation methods to obtain multi-step forecasts from the posterior distribution. For density forecasting, simulation methods work better than a direct multi-step approach, especially at long horizons. Specifications in levels benefit a lot from the imposition of the sum of coefficients and dummy initial observation priors of Doan *et al*. (1984) and Sims (1993). Instead, there is no payoff to using a Litterman (1986) prior that is tighter for lags of other variables than for lags of the dependent variable and treats the error variance matrix as fixed and diagonal. Using rolling estimation can enhance the density forecasting performance, while in terms of mean squared error it is difficult to do better than the benchmark. The finding that simple methods work well could therefore further enhance the diffusion of the BVAR as an econometric tool for a vast range of applications, particularly by researchers and practitioners relying on pre-programmed BVAR tools in common software packages such as RATS and Eviews.

ACKNOWLEDGEMENTS

We thank Frank Diebold, three anonymous referees, Ellis Tallman, Ken Beauchemin, Luc Bauwens and seminar participants at the Norges Bank for helpful comments, and Ethan Struby for research assistance. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

REFERENCES

Adolfson M, Linde J, Villani M. 2007. Forecasting Performance of an Open Economy DSGE Model. *Econometric Reviews* **26**: 289–328.

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

Banbura M, Giannone D, Reichlin L. 2010. Large Bayesian vector autoregressions. *Journal of Applied Econometrics* **25**: 71–92.

Bauwens L, Lubrano M, Richard JF. 1999. Bayesian Inference in Dynamic Econometric Models, Oxford University Press: Oxford.

Boivin J, Ng S. 2006. Are more data always better for factor analysis? *Journal of Econometrics* **132**: 169–194.

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

DOI: 10.1002/jae

Carriero A, Kapetanios G, Marcellino M. 2011. Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics* **26**: 735–761.
Carriero A, Kapetanios G, Marcellino M. 2012. Forecasting government bond yields with large Bayesian VARs. *Journal of Banking and Finance* **36**: 2026–2047.
Clark T. 2011. Real-time density forecasts from BVARs with stochastic volatility. *Journal of Business and Economic Statistics* **29**: 327–341.
Clark T, McCracken M. 2008. Forecasting with small macroeconomic VARs in the presence of instability. In Forecasting in the Presence of Structural Breaks and Model Uncertainty, Rapach D, Wohar M (eds). Emerald Group: Bingley, UK; 93–147.
Clark T, McCracken M. 2011a. Nested forecast model comparisons: a new approach to testing equal accuracy. Working paper, Federal Reserve Bank of St. Louis.
Clark T, McCracken M. 2011b. Testing for unconditional predictive ability. In Oxford Handbook of Economic Forecasting, Clements MP, Hendry DF (eds). Oxford University Press: Oxford; 415–440.
Clark T, McCracken M. 2011c. Advances in forecast evaluation. Working paper, Federal Reserve Bank of Cleveland.
Clements M, Hendry D. 1996. Intercept corrections and structural change. *Journal of Applied Econometrics* **11**: 475–494.
Del Negro M, Schorfheide F. 2004. Priors from general equilibrium models for VARs. *International Economic Review* **45**: 643–673.
Diebold FX, Kilian L. 2000. Unit-root tests are useful for selecting forecasting models. *Journal of Business and Economic Statistics* **18**: 265–273.
Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.
Doan T, Litterman R, Sims C. 1984. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews* **3**: 1–100.
Geweke J, Amisano G. 2010. Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, Elsevier **26**(2): 16–230.
Giannone D, Lenza M, Momferatou D, Onorante L. 2010. Short-term inflation projections: a Bayesian vector autoregressive approach. Working paper, European Central Bank.
Giannone D, Lenza M, Primiceri G. 2012. Prior selection for vector autoregressions. Working Paper No. 18467, NBER, Cambridge, MA.
Harvey D, Leybourne S, Newbold P. 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* **13**: 281–291.
Jacobson T, Karlsson S. 2004. Finding good predictors for inflation: a Bayesian model averaging approach. *Journal of Forecasting* **23**: 479–496.
Kadiyala K, Karlsson S. 1997. Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics* **12**: 99–132.
Karlsson S. 2012. Forecasting with Bayesian VAR models. In Handbook of Economic Forecasting, Vol. **2**. North-Holland: Amsterdam (forthcoming).
Koop G, Korobilis D. 2012. Large time-varying parameter VARs. Working paper, University of Strathclyde.
Koop G. 2013. Forecasting with medium and large Bayesian VARs. *J. Appl. Econ.* **28**: 177–203.
Leeper EM, Sims C, Zha T. 1996. What does monetary policy do? *Brookings Papers on Economic Activity* **27**: 1–78.
Litterman R. 1986. Forecasting with Bayesian vector autoregressions: five years of experience. *Journal of Business and Economic Statistics* **4**: 25–38.
Marcellino M, Stock J, Watson M. 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics* **127**: 499–526.
Osterholm P. 2008. Can forecasting performance be improved by considering the steady state? An application to Swedish inflation and interest rate. *Journal of Forecasting* **27**: 41–51.
Pesaran M, Pick A, Timmerman A. 2011. Variable selection and inference for multi-period forecasting problems. *Journal of Econometrics* **164**: 173–187.
Robertson JC, Tallman EW. 1999. Vector autoregressions: forecasting and reality. *Federal Reserve Bank of Atlanta Economic Review* First Quarter.
Sims C. 1993. A nine-variable probabilistic macroeconomic forecasting model. In Business Cycles, Indicators and Forecasting, Stock JH, Watson MW (eds). University of Chicago Press for NBER: Chicago, IL; 179–212.
Sims C, Zha T. 1998. Bayesian methods for dynamic multivariate models. *International Economic Review* **39**: 949–968.

DOI: 10.1002/jae

Waggoner DF, Zha T. 1999. Conditional forecasts in dynamic multivariate models. *The Review of Economics and Statistics* **81**: 639–651.
Wright J. 2009. Forecasting U.S. inflation by Bayesian model averaging. *Journal of Forecasting* **28**: 131–144.
Wright J. 2011. Evaluating real-time VAR forecasts with an informative democratic prior. *Journal of Applied Econometrics* DOI: 10.1002/jae.2268.
Zellner A. 1971. An Introduction to Bayesian Inference in Econometrics, Wiley: New York.
Zha T. 1998. A dynamic multivariate model for use in formulating policy. *Federal Reserve Bank of Atlanta Economic Review* First Quarter.