Evaluating the Accuracy of Forecasts from Vector Autoregressions

Evaluating the Accuracy of Forecasts from Vector Autoregressions *

Todd E. Clark
Federal Reserve Bank of Cleveland

Michael W. McCracken
Federal Reserve Bank of St. Louis

December 2012

Abstract

This paper surveys recent developments in the evaluation of point and density forecasts in the context of forecasts made by Vector Autoregressions. Specific emphasis is placed on highlighting those parts of the existing literature that are applicable to direct multi-step forecasts and those parts that are applicable to iterated multi-step forecasts. This literature includes advancements in the evaluation of forecasts in population (based on true, unknown model coefficients) and the evaluation of forecasts in the finite sample (based on estimated model coefficients). The paper then examines in Monte Carlo experiments the finite-sample properties of some tests of equal forecast accuracy, focusing on the comparison of VAR forecasts to AR forecasts. These experiments show the tests to behave as should be expected given the theory. For example, using critical values obtained by bootstrap methods, tests of equal accuracy in population have empirical size about equal to nominal size.

JEL Nos.: C53, C52, C12, C32
Keywords: prediction, forecasting, out-of-sample

* *Clark*: Economic Research Dept.; Federal Reserve Bank of Cleveland; P.O. Box 6387; Cleveland, OH 44101; todd.clark@clev.frb.org. *McCracken* (corresponding author): Research Division; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; michael.w.mccracken@stls.frb.org. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland, Federal Reserve Bank of St. Louis, Federal Reserve System, or any of its staff.

1 Introduction

Since the seminal work of Doan, Litterman, and Sims (1984), Litterman (1986), and Sims (1980), vector autoregressions (VARs) have become an essential tool for out-of-sample forecasting in macroeconomics. Their key role is evident in the long list of central bank research studies that deploy VARs, sometimes explicitly indicating VARs are in use at a particular central bank. While a complete listing is beyond the scope of the paper, recent examples of such studies include Coletti and Murchison (2002), Kapetanios, Labhard, and Price (2008), Bjornland, et al. (2012), Andersson, Karlsson, and Svensson (2007), Beauchemin and Zaman (2011), and Baumeister and Kilian (2012b,c). By now, VARs are commonly used for not only point forecasts but also density forecasts, sometimes in the form of fan charts.

Research on VAR forecasting has almost always included comparisons of the accuracy of VAR forecasts to the accuracy of forecasts from univariate benchmark models. Early work, such as Litterman (1986), commonly used random walk forecasts as a benchmark for comparison, reporting Theil's U statistic, without a formal measure of the statistical significance of differences in forecast accuracy. It has also become common to compare forecasts from VARs to forecasts from autoregressive models, since, for many (not all) variables of interest, an autoregressive model forecast is more accurate than a random walk. In recent work, it has become more common to provide some measure of the statistical significance of differences in forecast accuracy, sometimes with caveats about a need to interpret the results with some caution because existing asymptotic theory does not necessarily apply to forecasts from VARs. Some recent examples include Clark and McCracken (2010), Clark (2011), and Baumeister and Kilian (2012a). This trend of increased attempts to evaluate statistical significance highlights the fact that, while much is now known about proper inference for forecasts from a univariate model (summarized in surveys such as West (2006) and Clark and McCracken (2011b, 2012b)), much less is known about proper methods for testing equal accuracy of forecasts from VARs.

Accordingly, in this chapter, we review possible methods for evaluating the accuracy of out-of-sample forecasts from VARs. As we detail below, what methods are available (and appropriate) hinges on three aspects of the evaluation to be conducted. The first is the model specification used to produce multi-step forecasts. If the forecasts are generated with a direct multi-step model specification, existing results on the evaluation of forecasts from

a single model can be applied directly.[1] But if, as is more common for VARs, multi-step forecasts are produced by iterating forward 1-step ahead projections, some of the existing results either become much more complicated to use or cannot be used.

A second key determinant of available methods is the null hypothesis of interest – whether the relevant null is equal accuracy in the finite sample or equal accuracy in population.[2] The distinction between these two hypotheses is whether parameter estimation error is captured under the null hypothesis (finite sample) or whether the parameters are treated as known under the null (population). If the null of interest is the former, the results of Giacomini and White (2006) permit the use of t-tests of equal accuracy and standard normal critical values, but only for forecasts produced with a rolling window of data for model estimation and not for forecasts produced with an expanding window. Conceptually, this poses some challenge with the evaluation of VAR forecasts, since recursive estimation schemes are commonly used (although the rolling scheme is also regularly used). For a recursive scheme, at this point in time, there are no theoretically-justified methods for testing equal accuracy of VAR forecasts. From our Monte Carlo analysis, it seems as though the same tests that are theoretically appropriate under a rolling scheme and the null of equal accuracy in the finite sample may have reasonable size and power properties under a recursive scheme, but these tests lack a theoretical basis for the recursive scheme.

If the null hypothesis of interest is equal accuracy in population, the third key aspect of the forecasting design is the relationship of the VAR to the benchmark forecasting model. If the VAR does not nest the benchmark model, it is possible to rely on the results of West (1996) for conducting formal forecast inference, using standard asymptotic distributions. (If the forecasts are obtained by iteration, West's (1996) results apply, but can be computationally difficult.) However, if the VAR does nest the benchmark model of interest, as is often the case, inference becomes more complicated. At the 1-step horizon, the results of Clark and McCracken (2001, 2005a, 2012a) and McCracken (2007) provide a basis for inference, using asymptotic critical values obtained by Monte Carlo simulation or bootstrap. For forecasts at longer horizons obtained by iteration, it becomes difficult to derive the relevant asymptotic distributions. Still, in Monte Carlo experiments, some of the simulation

[1] Some might prefer that a system of equations all taking a direct multi-step form not be referred to as a vector autoregression. However, because other studies, such as Koop (2012) and Schorfheide (2005), have referred to these models as direct multi-step VARs, we follow their precedent.

[2] See Inoue and Kilian (2004) for an early example of a study drawing a distinction between equal accuracy in the finite sample and in population.

or bootstrap methods that work well at the 1-step horizon or with direct multi-step forecasts also seem to work well with iterated forecasts from VARs for the purpose of assessing predictability in population (as noted above, however, this finding does not pertain to the null of equal accuracy in the finite sample).

We begin the paper by detailing a brief summary of possible methods for testing equal accuracy of out-of-sample forecasts from VARs, focusing on the relevant econometric theory. Throughout our analysis we focus on evaluating accuracy variable-by-variable, rather than on evaluating accuracy for the system of variables as a whole. Although a few studies have considered multivariate measures of the accuracy of point forecasts (see, e.g., Komunjer and Owyang (2012)), most research on point forecasts focuses on univariate measures of accuracy. In the case of density evaluation, a measure of accuracy for the system of variables — the log predictive score for the vector of variables in the model — is perhaps more natural than in the case of point forecasts, but even studies of density accuracy commonly consider univariate measures. We leave as a subject for future research the further development of tests of the accuracy of point and density forecasts for a vector of variables.

We follow the theory treatment with a Monte Carlo analysis of a small range of tests of equal accuracy, using data generating processes that allow us to impose equal accuracy in population, but not in the finite sample. In the interest of brevity, we focus the Monte Carlo study on nested models, comparing VAR forecasts to AR forecasts, both point and density. In the Monte Carlo experiments, the tests of equal mean square error and forecast encompassing considered have size and power properties that seem reasonable. Perhaps most importantly, when the true model is a set of AR equations that are nested by the VAR model, and therefore the null hypothesis of no predictive accuracy in population is satisfied, using bootstrap methods to generate critical values yields rejection rates reasonably close to the nominal size of 10 percent. As would be expected and appropriate, tests of equal accuracy in the finite sample have lower rejection rates than tests of equal accuracy in population.

Throughout we emphasize tests of predictive ability based on forecasts generated by stationary VARs. We do so because the literature on tests of predictive ability is almost completely silent on inference when unit roots are present. The sole exception is Corradi, Swanson, and Olivetti (2001), who extend the analysis of West (1996) to permit cointegrating relationships but do so only for one-step ahead forecasts from univariate models. Their

results indicate that in certain special cases, tests of predictive ability are asymptotically normal as in West (1996). More generally, these tests have non-standard asymptotic distributions due to the presence of unit roots. They argue that many of their results extend directly to longer horizon, direct multi-step forecasts but do not show this explicitly. Regardless, there exist no results that are specifically applicable to VARs with unit roots and hence we focus our attention on stationary VARs.

In addition, we focus exclusively on analytical results that ignore real-time vintage data issues. We do so in order to focus attention on the primary results within the literature without having to introduce the additional notation needed to keep track of data revisions. This should not be interpreted as meaning that real-time data issues are unimportant for out-of-sample inference. To the contrary, Clark and McCracken (2009) extend the results in West (1996) allowing for real-time data and find that while many test statistics remain asymptotically normal they do so with a distinct asymptotic variance structure due to the presence of data revisions. In the context of this paper, the topic is of particular importance since VARs are a popular tool for forecasting at many central banks which, by their very nature, must construct forecasts in real-time using sequences of vintage data. Even so, for clarity we omit the issue within the remainder.

The rest of the paper proceeds as follows. Section 2 gives a very brief overview of estimation and iterated multi-step forecasting with stationary VARs, with an emphasis on how these forecasts differ from those based on direct multi-step models. Section 3 discusses the existing literature in the context of tests of population predictive ability as compared to tests of finite sample predictive ability. Section 4 presents Monte Carlo results on the performance of some basic testing procedures. Section 5 concludes.

2 Constructing the Forecasts

Before examining options for testing the predictive ability of VAR-based forecasts, it is helpful to review methods for generating the forecasts. Researchers and practitioners use a range of approaches to estimating VAR models and generating forecasts. As we explain below, how the forecasts are generated bears on how the forecasts can be evaluated. This section proceeds by first distinguishing iterated multi-step (IMS) and direct multi-step (DMS) approaches to producing forecasts and then detailing the most common VAR estimation approaches.

2.1 DMS vs. IMS Forecasts

In the VAR forecasting literature, it is most common to generate multi-step point forecasts with the IMS approach (detailed below). However, it is sometimes the case that multi-step point forecasts are obtained with a DMS approach (e.g., Koop 2012). As Schorfheide (2005) and Marcellino, Stock, and Watson (2006) note, the DMS approach may yield more accurate forecasts than the IMS approach if the VAR is sufficiently misspecified. In the VAR results of Carriero, Clark, and Marcellino (2012), IMS and DMS forecast accuracy is comparable, with a little advantage to the IMS approach, more so for density forecasts than point forecasts.

While IMS forecasts are more common than DMS forecasts in VAR-based work, the testing literature tends to, either theoretically or by example, focus on procedures designed to accommodate forecasts constructed using the DMS approach. In some parts of the literature, like that of Clark and McCracken (2001, 2005a), theoretical results on tests of equal forecast accuracy between two nested models only apply to DMS forecasts. In other parts of the literature, like West (1996) or McCracken (2000), both IMS and DMS forecasts are allowed but the theory is much more easily implemented when DMS rather than IMS forecasts are used and hence the empirical applications focus on DMS forecasts.

To get a feel for the difference between the two forecasting procedures and for why results for out-of-sample tests of predictive ability may differ when IMS vs. DMS forecasts are used, suppose we are interested in forecasting a scalar y_1 variable τ-periods in the future using the current, time t value of y_1 as a predictor. A DMS approach to point forecasting would use OLS to estimate the parameters in a simple linear regression of the form

$$y_{1,t} = \beta_0 + \beta_1 y_{1,t-\tau} + \varepsilon_{1,t} = \beta' x_{t-\tau} + \varepsilon_{1,t} \tag{1}$$

and then construct the point forecast as $\hat{y}_{1,t+\tau} = \hat{\beta}_{0,t} + \hat{\beta}_{1,t} y_{1,t} = \hat{\beta}_t' x_t$ where $\hat{\beta}_t = (\hat{\beta}_{0,t}, \hat{\beta}_{1,t})' = (t^{-1}\sum_{s=1}^{t-\tau} x_s x_s')^{-1}(t^{-1}\sum_{s=1}^{t-\tau} x_s y_{1,s+\tau})$. In contrast an IMS approach to forecasting would use OLS to estimate the parameters in a slightly different simple linear regression model of the form

$$y_{1,t} = \beta_0 + \beta_1 y_{1,t-1} + \varepsilon_{1,t} = \beta' x_{t-1} + \varepsilon_{1,t}. \tag{2}$$

With these parameters in hand, and with an implicit assumption that the errors in equation (2) form a martingale difference sequence, the IMS point forecast of $y_{1,t+\tau}$ is $\hat{y}_{1,t+\tau} = \hat{\beta}_{0,t}(1+$

$\sum_{j=1}^{\tau-1} \hat{\beta}_{1,t}^{j}) + \hat{\beta}_{1,t}^{\tau} y_{1,t}$, where $\hat{\beta}_t = (\hat{\beta}_{0,t}, \hat{\beta}_{1,t})' = (t^{-1} \sum_{s=1}^{t-1} x_s x_s')^{-1} (t^{-1} \sum_{s=1}^{t-1} x_s y_{1,s+1})$. The DMS and IMS forecasts are clearly related to one another. They are both linear functions of $x_t = (1, y_{1,t})'$. Moreover, if the errors in equation (2) form a martingale difference sequence, it is straightforward to show that (the population parameters) β_0 and β_1 in equation (1) equal $\beta_0(1 + \sum_{j=1}^{\tau-1} \beta_1^j)$ and β_1^τ in equation (2) and hence in large enough samples the forecasts are identical.

That said, as a practical matter the point forecasts made using the IMS approach are a more complicated function of the parameter estimates $\hat{\beta}_t$ than they are when using the DMS approach. The complexity of the problem only increases as we move from IMS forecasts from autoregressive (AR) models to IMS forecasts from VARs. To see this consider extending the models in (1) and (2) to allow information from other time t predictors. In particular suppose we are interested in forecasting a scalar y_1 variable τ-periods in the future using the current, time t value of y_1 as a predictor and the time t values of predictors y_2 and y_3. A DMS approach to point forecasting would use OLS to estimate the parameters in the linear regression

$$y_{1,t} = \beta_0 + \beta_1 y_{1,t-\tau} + \beta_2 y_{2,t-\tau} + \beta_3 y_{3,t-\tau} + \varepsilon_{1,t} = \beta' x_{t-\tau} + \varepsilon_{1,t} \tag{3}$$

and then construct the point forecast as $\hat{y}_{1,t+\tau} = \hat{\beta}_{0,t} + \hat{\beta}_{1,t} y_{1,t} + \hat{\beta}_{2,t} y_{2,t} + \hat{\beta}_{3,t} y_{3,t} = \hat{\beta}_t' x_t$ where $x_t = (1, y_{1,t}, y_{2,t}, y_{3,t})'$ and $\hat{\beta}_t = (t^{-1} \sum_{s=1}^{t-\tau} x_s x_s')^{-1} (t^{-1} \sum_{s=1}^{t-\tau} x_s y_{1,s+\tau})$. In contrast an IMS approach to forecasting would use OLS to estimate the parameters in the three equation linear system

$$\begin{aligned} \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} &= \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} + \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{pmatrix} \\ &= C + A Y_{t-1} + \varepsilon_t, \end{aligned} \tag{4}$$

which is equivalent to

$$\begin{aligned} Y_t &= \Lambda x_{t-1} + \varepsilon_t \\ &= (x_{t-1}' \otimes I_3)\beta + \varepsilon_t, \end{aligned}$$

if we define $Y_t = (y_{1,t}, y_{2,t}, y_{3,t})'$, $\beta = vec(\Lambda)$, and

$$\Lambda = \begin{pmatrix} c_1 & A_{11} & A_{12} & A_{13} \\ c_2 & A_{21} & A_{22} & A_{23} \\ c_3 & A_{31} & A_{32} & A_{33} \end{pmatrix}.$$

With these parameters in hand, and with an implicit assumption that the errors in equation (4) form a martingale difference sequence, the IMS point forecast of $y_{1,t+\tau}$ is $\hat{y}_{1,t+\tau} = \iota'((I+\sum_{j=0}^{\tau-1}\hat{A}_t^j)\hat{C}_t+\hat{A}_t^\tau Y_t)$ where $\iota' = (1,0,0)$ and $\hat{\Lambda}_t = (t^{-1}\sum_{s=1}^{t-1} Y_{s+1}x_s')(t^{-1}\sum_{s=1}^{t-1} x_s x_s')^{-1}$. The DMS and IMS forecasts continue to be related to one another. They are both linear functions of x_t and as was the case above, if the errors in equation (4) form a martingale difference, it is straightforward to show that (the population) β in equation (3) equals $(\iota'(I + \sum_{j=0}^{\tau-1} A^j)C, \iota' A^\tau)$ and hence in large enough samples the forecasts are identical. Regardless, the IMS forecasts remain a much more complicated function of β for the IMS forecasts in (4) than for the DMS forecasts in (3).

2.2 Estimating and forecasting with VARs

Of course the comparison between DMS and IMS forecasts can be generalized further to allow more lags and more predictors in the system. Accordingly, it will be useful to generalize the notation for VAR-based IMS forecasts for use throughout the remainder of the text. To do so suppose that y_1 is the scalar variable we wish to predict at the τ-step horizon but now assume that there exist $n-1$ other predictors and that we allow for $m \geq 1$ lags. In lag order notation the VAR model takes the form

$$Y_t = C + A(L)Y_{t-1} + \varepsilon_t,$$

where $Y = (y_1, y_2, ..., y_n)'$, $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ and $A(L) = \sum_{j=1}^m A_j L^j$ for $n \times 1$ and $n \times n$ parameter matrices C and A_j, $j = 1, ..., m$, respectively. This is again equivalent to

$$\begin{aligned} Y_t &= \Lambda x_{t-1} + u_t \\ &= (x_{t-1}' \otimes I_n)\beta + \varepsilon_t, \end{aligned}$$

if we define $x_t = (1, Y_t', ..., Y_{t-m+1}')'$, $\beta = vec(\Lambda)$, and

$$\Lambda = (C, A_1, ..., A_m).$$

Under the maintained assumption that the model errors ε_t form a martingale difference sequence, the IMS τ-step ahead forecasting function $\hat{y}_{1,t+\tau} = \iota'\hat{Y}_{t+\tau}$, with $\iota = (1, 0, ..., 0)'$, can be inferred from the recursion $\hat{Y}_{t+\tau} = \iota'(C + A_1\hat{Y}_{t+\tau-1} + ... + A_m\hat{Y}_{t+\tau-m})$, where for the moment we abstract from how the regression parameters are estimated. But of course the regression parameters do need to be estimated before the forecasts can be constructed.

While there are many ways to do so, we focus on OLS and Bayesian estimation since they appear to be the most common in current applied work.

2.2.1 OLS Estimation

As noted in the previous section, the OLS estimates of Λ take the simple closed form $\hat{\Lambda}_t = (t^{-1}\sum_{s=1}^{t-1} Y_{s+1}x_s')(t^{-1}\sum_{s=1}^{t-1} x_s x_s')^{-1}$. For reference in the following section it is worth noting that if we're willing to assume that ε_t forms a martingale difference sequence and the observables are stationary, the vector of parameter estimates can be represented as equivalent to the true parameter values plus a term driven by estimation error:

$$\hat{\Lambda}_t = \Lambda + (t^{-1}\sum\nolimits_{s=1}^{t-1} \varepsilon_{s+1}x_s')(t^{-1}\sum\nolimits_{s=1}^{t-1} x_s x_s')^{-1}.$$

Equivalently, if we map this into the definition of $\beta = vec(\Lambda)$, this takes the form

$$\hat{\beta}_t = \beta + B(t)H(t),$$

where $B = ((t^{-1}\sum_{s=1}^{t-1} x_s x_s')^{-1} \otimes I_n)$ and $H(t) = vec(t^{-1}\sum_{s=1}^{t-1} \varepsilon_{s+1}x_s')$.

2.2.2 Bayesian Estimation

By now, in the VAR forecasting literature, it is probably more common to use models estimated by Bayesian methods than models estimated by least squares, particularly with larger models. One key reason is that studies such as Clark and McCracken (2008), Banbura, Giannone, and Reichlin (2010), and Wright (2012) have shown forecasts from BVARs to often be more accurate than forecasts from OLS-estimated VARs, with gains that increase in the size of the model (however, there are examples, such as Baumeister and Kilian (2012a), in which an OLS-estimated VAR forecasts better than a BVAR). In no small part this is due to the number of degrees of freedom used up when estimating a VAR. Bayesian methods, like ridge regression techniques, use shrinkage of the parameters towards a given set of priors in order to increase the (finite-sample) precision of the parameters.

As detailed in resources such as Kadiyala and Karlsson (1997), Geweke and Whiteman (2006), and Karlsson (2012), BVARs can be estimated with a wide array of approaches, depending in part on the priors that are used.[3] Due to computational constraints, Litterman (1986) developed an approach that is equivalent to applying ridge regression on an equation-by-equation basis, based on an assumption of an error covariance matrix that is fixed and

[3]Giannone, Lenza, and Primiceri (2012) extend the common Normal-Wishart setup to incorporate a hierarchical prior that covers the usual hyperparameters of the Minnesota prior.

diagonal. With current computational power, Monte Carlo simulation methods are now often seen as more or less the standard for BVAR forecasting (except with large models). The simulation method required hinges on the particular prior used. We will proceed to briefly describe some approaches and then conclude with a discussion of how Bayesian estimation fits within the methods admitted in the theory underlying forecast evaluation.

Among possible prior specifications or approaches, the simplest case is the Normal-Wishart prior, which yields a posterior of the same form. Under a Normal-Wishart prior, the posterior mean of the VAR coefficients can be computed directly, without simulation. As a consequence, the posterior mean of 1-step ahead forecasts, which is a linear function of the VAR coefficients, can also be computed directly, without simulation. Studies such as Carriero, Clark, and Marcellino (2012) and Koop (2012) use a Normal-Wishart prior with VARs specified in DMS form to similarly obtain multi-step forecasts without simulation.[4] In the IMS case, because the multi-step forecasts are non-linear functions of the VAR coefficients, computing the mean of the posterior distribution of forecasts requires simulation of the distribution. In practice, though, Carriero, Clark, and Marcellino (2012) find that the properly simulated mean is essentially the same as an approximate mean computed by using the posterior mean coefficients to compute multi-step forecasts by iteration, as outlined above.

More specifically, consider the implementation of a Normal-Wishart prior and posterior, abstracting for now from the details of the prior. By grouping the coefficients and explanatory variables, the VAR can be written as

$$Y_t = \Lambda x_{t-1} + \varepsilon_t, \tag{5}$$

or, even more compactly, as:

$$Y = X\Lambda' + E, \tag{6}$$

where $Y = [Y_1, .., Y_T]'$, $X = [x_1, .., x_T]'$, and $E = [\varepsilon_1, .., \varepsilon_T]'$ are, respectively, $T \times n$, $T \times (nm + 1)$ and $T \times n$ matrices. The conjugate Normal-Wishart prior takes the form:

$$\Lambda'|\Sigma \sim N(\Lambda_0', \Sigma \otimes \Omega_0),\ \Sigma \sim IW(S_0, v_0), \tag{7}$$

which yields a conditional posterior distribution that is also Normal-Wishart (Zellner 1971):

$$\Lambda'|\Sigma, Y \sim N(\bar{\Lambda}', \Sigma \otimes \bar{\Omega}),\ \Sigma|Y \sim IW(\bar{S}, \bar{v}). \tag{8}$$

[4]However, such an approach ignores the serial correlation in the errors of a VAR in DMS form.

Defining $\hat{\Lambda}'$ and $\hat{E}$ as the OLS estimates, we have that

$$\begin{aligned}
\bar{\Lambda}' &= (\Omega_0^{-1} + X'X)^{-1}(\Omega_0^{-1}\Lambda_0' + X'Y) \\
\bar{\Omega} &= (\Omega_0^{-1} + X'X)^{-1} \\
\bar{v} &= v_0 + T \\
\bar{S} &= \Lambda_0' + \hat{E}'\hat{E} + \hat{\Lambda}X'X\hat{\Lambda}' + \Lambda_0\Omega_0^{-1}\Lambda_0' - \bar{\Lambda}\bar{\Omega}^{-1}\bar{\Lambda}'.
\end{aligned} \tag{9}$$

Karlsson (2012) details the steps involved in simulating the posterior distribution.

This Normal-Wishart approach has both advantages and disadvantages. One advantage is that the posterior mean of the coefficients can be computed without simulation, which means that 1-step ahead forecasts can also be computed without simulation and multi-step point forecasts can at least be approximated very well without simulation. These considerations have led to the use of the Normal-Wishart prior and posterior (without simulation) in large model studies such as Banbura, Giannone, and Reichlin (2010) and Koop (2012). A second advantage is that the posterior distribution can be quickly and easily simulated. The simulations are quick because the posterior variance of the coefficients has a Kronecker structure, which greatly speeds the necessary matrix computations. The simulations are easy because the distributions can be directly simulated, without resorting to full Markov chain methods such as Gibbs sampling. The disadvantage of the Normal-Wishart approach is that it does not permit one of the key components of the Minnesota prior as developed by Litterman (1986): it does not permit one to make the prior on lags of variable j in equation i tighter than the prior on lags of variable i in the same equation. The Normal-Wishart specification instead requires the prior variances to be symmetric across equations.

Avoiding this disadvantage of the Normal-Wishart specification requires using Markov chain methods such as Gibbs sampling to estimate the BVAR. For example, to implement a Minnesota prior that includes a different rate of shrinkage on "own" lags versus "other" lags, the most common approach rests on the Normal-diffuse prior detailed in Kadiyala and Karlsson (1997), among others. The model is then estimated by Gibbs sampling, using conditional distributions for the VAR coefficients and the error variance-covariance matrix. Karlsson (2012) details the steps involved in this sampling. While readily tractable with smaller models, the computations become more costly as the model size increases, due to the need for repeated computations with a large variance-covariance matrix of coefficients

that lacks a Kronecker structure.[5]

Regardless of the particular approach used to estimate a VAR by Bayesian methods, Bayesian estimation can be viewed as falling within the scheme of model estimation approaches assumed in the evaluation theory reviewed in the next section. Most immediately, because the asymptotic approach of Giacomini and White (2006) does not involve any assumptions about model estimation, forecasts from Bayesian VARs are allowed in their framework. While the asymptotic approaches of West (1996), West and McCracken (1998), and Clark and McCracken (2001, 2005a) each make assumptions about model estimation methods, Bayesian estimation can be seen as fitting within these assumptions. Consider, for example, the nested model theory of Clark and McCracken (2001, 2005a), which requires some form of least squares estimation, but uses asymptotics that treat the estimation sample size as diverging to infinity. As the sample size used for estimation expands, the importance of the prior used in Bayesian estimation declines, reaching zero in the limit. This is clear in the posterior mean of the VAR coefficients given in equation (9). As the sample size grows to infinity, the data terms $X'X$ and $X'Y$ will overwhelm the prior terms Ω_0 and $\Omega_0^{-1}\Lambda_0$, and the posterior mean of the VAR coefficients will converge to the OLS estimates and, in turn, the population value. Therefore, asymptotically, Bayesian estimation can be seen as equivalent to the least squares estimation admitted by the asymptotic theory. Admittedly, this interpretation is not as straightforward under priors that require Gibbs sampling or other alternatives to the Normal-Wishart prior and posterior. But even in these cases, it will generally be the case that, asymptotically, the posterior mean of the VAR coefficient estimates will converge to the OLS estimates.

3 Theory

In this section we provide an overview of asymptotically valid tests of predictability with an eye towards distinguishing between inference when DMS or IMS forecasts are being used. For each test, the statistics are sample averages of functions of a sequence of forecasts and other observables available at the time of the forecast or when the forecast error is observed.

[5]Carriero, Clark, and Marcellino (2012) investigate the performance of a Litterman-style approach that is more tractable for simulation because it treats each equation separately, under the assumption of a fixed and diagonal error covariance matrix. They don't find such an approach to improve on a Normal-Wishart setup, partly because the negative consequences of the assumption of a fixed and diagonal error covariance matrix offset the positive effects of imposing cross-variable shrinkage.

3.1 Examples and Notation

Before proceeding to the theoretical results we lay out some notation in the context of the example of a test of zero mean prediction error when forecasts are made at a four-step ahead horizon. We do so once assuming DMS forecasts akin to that in equation (3) and once assuming IMS forecasts using a quarterly frequency trivariate VAR(1) akin to that in equation (4). We then generalize this very simple testing procedure to pairwise tests of equally accurate point and density forecasts.

Suppose we have quarterly data from 1947:Q1 - 2012:Q2 for a total of $T = 262$ observations. The sample is split into an in-sample period and an out-of-sample period, respectively. The former is used to estimate the parameters associated with the first forecast while the latter denotes the number of forecast origins from which four quarter-ahead ($\tau = 4$) forecasts can be constructed. Let 1947:Q1 - 1979:Q4 denote the in-sample period and 1980:Q1 - 2011:Q2 denote the out-of-sample period, giving us $R = 132$ in-sample observations and $P - 3 = 127$ out-of-sample periods. Note that data from 2011:Q3 - 2012:Q2 are not treated as forecast origins but will be used to construct the forecast errors.

How we construct the test of zero mean prediction error depends on whether DMS or IMS forecasts are being used.

1. DMS forecast: At each forecast origin $t = R, ..., T - 4$, the linear regression model in equation (3) is estimated by OLS using all available data and hence
$\hat{\beta}_t = (t^{-1} \sum_{s=1}^{t-4} x_s x_s')^{-1} (t^{-1} \sum_{s=1}^{t-4} x_s y_{1,s+4})$. Together with the predictors x_t these parameter estimates yield a sequence of forecast errors of the form $\hat{\varepsilon}_{1,t+4} = y_{1,t+4} - \hat{y}_{1,t+4}$, where $\hat{y}_{1,t+4} = x_t' \hat{\beta}_t$. The test of zero mean prediction error is based on a statistic of the form $(P-3)^{-1/2} \sum_{t=R}^{T-4} \hat{\varepsilon}_{1,t+4} / \hat{\Omega}^{1/2}$, where $\hat{\Omega}$ is a consistent estimate of an unknown long run variance Ω (characterized later in section 3.2). Note that the numerator of the statistic can be written in the generic form $(P-3)^{-1/2} \sum_{t=R}^{T-4} f(X_{t+4}, \hat{\beta}_t)$, where X_{t+4} is the vector of observables $(y_{1,t+4}, x_t')'$ and $f(X_{t+4}, \hat{\beta}_t) = \hat{\varepsilon}_{1,t+4}$. In the next section we describe how to consistently estimate Ω and conduct inference.

2. IMS forecast: At each forecast origin $t = R, ..., T - 4$, the VAR(1) model in equation (4) is estimated by OLS using all available data and hence
$\hat{\Lambda}_t = (t^{-1} \sum_{s=1}^{t-1} Y_{s+1} x_s')(t^{-1} \sum_{s=1}^{t-1} x_s x_s')^{-1}$. With these parameter estimates in hand, forecasts of $y_{1,t+4}$ take the form $\hat{y}_{1,t+4} = \iota'((I + \sum_{j=0}^{3} \hat{A}_t^j) \hat{C}_t + \hat{A}_t^4 Y_t)) = \iota'((I + \sum_{j=0}^{3} \hat{A}_t^j) \hat{C}_t, \hat{A}_t^4) x_t$, with subsequent forecast errors $\hat{\varepsilon}_{1,t+4} = y_{1,t+4} - \hat{y}_{1,t+4}$. Based on these forecast er-

rors a test of zero mean prediction error is constructed using the same statistic $(P-3)^{-1/2}\sum_{t=R}^{T-4}\hat{\varepsilon}_{1,t+4}/\hat{\Omega}^{1/2}$, where $\hat{\Omega}$ is a consistent estimate of an unknown long run variance Ω which may differ from that above associated with DMS forecasts. As was the case for the DMS forecast above, the numerator of the statistic can be written in the generic form $(P-3)^{-1/2}\sum_{t=R}^{T-4} f(X_{t+4},\hat{\beta}_t)$, where X_{t+4} is the vector of observables $(y_{1,t+4},x_t')'$ and $f(X_{t+4},\hat{\beta}_t)=\hat{\varepsilon}_{1,t+4}$, but only after defining $\hat{\beta}_t=vec(\hat{\Lambda}_t)$. It's also worth noting that since $\hat{\varepsilon}_{1,t+4}$ is a fourth order polynomial in the elements of $\hat{\Lambda}_t$, the function $f(X_{t+4},\hat{\beta}_t)$ is a far more complex function of $\hat{\beta}_t$ when IMS forecasts are used than it is when DMS forecasts are used.

This simple example of testing for zero mean prediction error is easily generalized to a wider range of null hypotheses of the form $H_0: Ef(X_{t+4},\beta^*)=0$. Perhaps the most common of these tests are ones designed to compare the relative accuracy of two models. For example, suppose that in the zero mean prediction error example we actually had two models, indexed as $i=1,2$, that were used to construct forecasts (either both DMS or both IMS) rather than just one model. Both models could be used to construct a sequence of forecast errors $\hat{\varepsilon}_{1,t+4}^{(i)}$, $t=R,...,T-4$, and, for a given loss function $L(\cdot)$, the relative accuracy of the two models could be evaluated. By far the most common of these loss functions is the quadratic, for which $L(\varepsilon)=\varepsilon^2$. If this loss function is used the relative accuracy of the two models can be tested using the same framework described above if we define $f(X_{t+4},\hat{\beta}_t)=(\hat{\varepsilon}_{1,t+4}^{(1)})^2-(\hat{\varepsilon}_{1,t+4}^{(2)})^2$, where X_{t+4} denotes the vector of observables used to construct both forecast errors and $\hat{\beta}_t$ is defined as the stacked vector of parameter estimates $\hat{\beta}_t=(\hat{\beta}_t^{(1)\prime},\hat{\beta}_t^{(2)\prime})'$ from the two models.

While quadratic loss is the most common among all loss functions, density forecasts are increasingly common. Most typically, the density forecasts are obtained by Monte Carlo simulation, using an IMS approach to horizons greater than one. However, studies such as Koop (2012) obtain density forecasts directly (i.e., without simulation) from the t-distribution implied by a Normal-Wishart posterior and a DMS specification. Regardless, suppose that two distinct VARs are used, each of which implies a distinct predictive density $g_i(X_{t+\tau},\hat{\beta}_t^{(i)})$. Amisano and Giacomini (2007) consider comparing the accuracy of the two VARs based on their relative log predictive scores. In our notation, this is equivalent to testing the (population) null hypothesis that $H_0: E[\ln g_1(X_{t+\tau},\beta^{(1)*})-\ln g_2(X_{t+\tau},\beta^{(2)*})]=0$ using the function $f(X_{t+\tau},\hat{\beta}_t)=\ln g_1(X_{t+\tau},\hat{\beta}_t^{(1)})-\ln g_2(X_{t+\tau},\hat{\beta}_t^{(2)})$ to construct the test

statistic $(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t)/\hat{\Omega}^{1/2}$.

3.2 Population tests of predictive ability

In reviewing existing theory for VAR forecast evaluation, we begin with inference based on West (1996), which, for comparisons of models, can be applied with non-nested models only. In subsequent sections we take up nested model comparisons and then testing based on Giacomini and White (2006) asymptotics.

Regardless of whether DMS or IMS forecasts are used, the theory in West (1996) can be used to conduct asymptotically valid inference for many (though not all) statistics that are scaled sample averages of the form

$$(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t) \tag{10}$$

under modest assumptions on (i) the observables $X_{t+\tau}$, (ii) how the parameter estimates $\hat{\beta}_t$ are constructed, and (iii) smoothness conditions on the function $f(X_{t+\tau},\beta)$ as a function of β. The major building block of this theory is Lemma 4.3 of West (1996), in which it is shown that under the null hypothesis $H_0: Ef(X_{t+\tau},\beta^*)=0$, the out of sample average of the function $f(X_{t+\tau},\hat{\beta}_t)$ can be decomposed into two parts:

$$\begin{aligned} & (P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t) \qquad (11)\\ = \;& (P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\beta^*) + FB((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} H(t)) + o_p(1), \end{aligned}$$

where $F = E\partial f(X_{t+\tau},\beta)/\partial\beta|_{\beta=\beta^*}$, B is a non-stochastic matrix satisfying $\hat{\beta}_t-\beta^* = BH(t)+o_{a.s.}(1)$, and $H(t)=t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$ for a sequence of mean zero increments $h_{s+\tau}$. The first right-hand side term in equation (11) denotes that part of the statistic associated with the forecast had the population values of the parameters been known while the second right-hand side term denotes that part of the statistic associated with the fact that the model parameters cannot be observed but instead must be estimated.

Building on Lemma 4.3, West (1996) shows that the statistic $(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t)/\hat{\Omega}^{1/2}$ is asymptotically standard normal so long as Ω is positive and $\hat{\Omega}$ is consistent for its population counterpart. The details of how to estimate Ω are perhaps the main technical developments in West (1996). Before providing this result, some additional notation and assumptions are needed.[6]

[6]These assumptions are intended to be expository, not complete. See West (1996) for more detail.

(A1) $\hat{\beta}_t = \beta^* + BH(t) + o_{a.s.}(1)$, where for some mean zero process $h_{t+\tau} = h_{t+\tau}(\beta^*)$ [with h denoting the orthogonality conditions used to estimate parameters, such as $h_{t+\tau} = x_t u_{t+\tau}$ for a single linear regression], $H(t)$ equals $t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$, $R^{-1}\sum_{s=t-R+1}^{t-\tau} h_{s+\tau}$, and $R^{-1}\sum_{s=1}^{R-\tau} h_{s+\tau}$ for the recursive, rolling, and fixed schemes, respectively, and B denotes a non-stochastic matrix.

(A2) The vector $(f(X_{t+\tau}, \beta^*), h'_{t+\tau})'$ is covariance stationary and satisfies mild mixing and moment conditions.

(A3) $\lim_{P,R\to\infty} P/R = \pi$, a constant that is finite for the rolling and fixed schemes but can be infinite for the recursive scheme.

(A4) The vector $F = E[\partial f(X_{t+\tau}, \beta)/\partial\beta]_{\beta=\beta^*}$ is finite.[7]

(A5) Ω is positive definite.

Given these assumptions, West (1996) shows that the asymptotic variance Ω can take a variety of forms depending on how the parameters are estimated:

$$\Omega = S_{ff} + 2\lambda_{fh} FBS'_{fh} + \lambda_{hh} FBS_{hh} B'F', \tag{12}$$

where $S_{ff} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} f(X_{s+\tau}, \beta^*))$, $S_{hh} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} h_{s+\tau})$, $S_{fh} = \lim_{T\to\infty} Cov(T^{-1/2}\sum_{s=1}^{T-\tau} f(X_{s+\tau}, \beta^*), T^{-1/2}\sum_{s=1}^{T-\tau} h_{s+\tau})$, and

	$\lambda_{fh} =$	$\lambda_{hh} =$
Recursive	$1 - \pi^{-1}\ln(1+\pi)$	$2(1 - \pi^{-1}\ln(1+\pi))$
Rolling, $\pi \leq 1$	$\pi/2$	$\pi - \pi^2/3$
Rolling, $1 < \pi < \infty$	$1 - (2\pi)^{-1}$	$1 - (3\pi)^{-1}$
Fixed	0	π

.

In equation (12) we see that Ω consists of three terms. The first, S_{ff}, is the long-run variance of the measure of accuracy when the parameters are known. The third term, $\lambda_{hh} FBS_{hh} B'F'$, captures the contribution of the variance due purely to the fact that we do not observe β^* but must estimate it instead. The second term, $2\lambda_{fh} FBS'_{fh}$, captures the covariance between the measure of accuracy and the estimation error associated with $\hat{\beta}_t$. Because the parameter estimates can be constructed using three different observation windows (recursive, rolling, and fixed) it is not surprising that the terms that arise due to estimation error depend on that choice via the terms λ_{fh} and λ_{hh}.

Estimating Ω is reasonably straightforward when either IMS or DMS forecasts are used but can be considerably simpler when DMS forecasts are used. Regardless of which approach is taken, since $\hat{\pi} = P/R \to \pi$ and both λ_{fh} and λ_{hh} are continuous in π, substituting

[7] McCracken (2000) weakens this assumption to $F = \partial E[f(X_{t+\tau}, \beta)]/\partial\beta_{\beta=\beta^*}$ so that the function $f(X_{t+\tau}, \beta)$ need not be differentiable in β.

$\hat{\pi}$ for π is sufficient for estimating both λ_{fh} and λ_{hh}. For the remaining components one needs to know whether the IMS or DMS approach is used to construct multi-step forecasts. The distinction matters for how B, $h_{t+\tau}$, and F are defined and, in turn, how each is to be estimated.

To see this consider the test of zero mean prediction error described above with $f(X_{t+4}, \hat{\beta}_t) = \hat{\varepsilon}_{1,t+4}$. In the following we delineate how each of the remaining components of Ω can be estimated depending on whether DMS or IMS forecasts are used.

- B: For the DMS case $B = (Ex_t x_t')^{-1}$ and hence $\hat{B} = (T^{-1}\sum_{s=1}^{T-4} x_s x_s')^{-1}$ is a consistent estimator of B. For the IMS case $B = ((Ex_t x_t')^{-1} \otimes I_n)$ and hence $\hat{B} = ((T^{-1}\sum_{s=1}^{T-4} x_s x_s')^{-1} \otimes I_n)$ is a consistent estimator of B.[8]

- S_{ff}, S_{fh}, and S_{hh}: For the long-run variances and covariances needed to compute the test statistic, West (1996) shows that standard kernel-based estimators, which are averages of $f(X_{t+4}, \hat{\beta}_t)$ and $h_{t+4}(\hat{\beta}_t)$, are consistent. We know $f(X_{t+4}, \hat{\beta}_t) = \hat{\varepsilon}_{1,t+4}$ for both the IMS and DMS cases. For the DMS case it suffices to define $h_{t+4}(\hat{\beta}_t) = x_t \hat{\varepsilon}_{1,t+4}$ while for the IMS case it suffices to define $h_{t+4}(\hat{\beta}_t) = vec(\hat{\varepsilon}_{t+4} x_t')$. With these in hand if we define $\bar{f} = (P-3)^{-1}\sum_{t=R}^{T-4} f(X_{t+4}, \hat{\beta}_t)$, $\hat{\Gamma}_{ff}(j) = (P-3)^{-1}\sum_{t=R+j}^{T-\tau}(f(X_{t+4}, \hat{\beta}_t) - \bar{f})(f(X_{t+4-j}, \hat{\beta}_{t-j}) - \bar{f})'$, $\hat{\Gamma}_{hh}(j) = T^{-1}\sum_{t=j+1}^{T-4} h_{t+4}(\hat{\beta}_t) h_{t+4-j}'(\hat{\beta}_{t-j})$ and $\hat{\Gamma}_{fh}(j) = (P-3)^{-1}\sum_{t=R+j}^{T-\tau} f(X_{t+4}, \hat{\beta}_t) h_{t+4-j}'(\hat{\beta}_{t-j})$, with $\hat{\Gamma}_{ff}(j) = \hat{\Gamma}_{ff}(-j)$, $\hat{\Gamma}_{hh}(j) = \hat{\Gamma}_{hh}'(-j)$, and $\hat{\Gamma}_{fh}(j) = \hat{\Gamma}_{fh}'(-j)$, the long-run variance estimates $\hat{S}_{ff}$, $\hat{S}_{hh}$, and $\hat{S}_{fh}$ are then constructed by weighting the relevant leads and lags of these covariances, as in HAC estimators such as the one developed by Newey and West (1987).[9]

- F: This term poses probably the greatest complication in estimating Ω, particularly with the IMS approach to generating multi-step forecasts. To see this recall that $F = E[\partial f(X_{t+4}, \beta)/\partial\beta]_{\beta=\beta^*}$.[10] In the DMS case this is trivially equal to $-Ex_t$ and can easily be estimated using $\hat{F} = -(P-3)^{-1}\sum_{t=R}^{T-4} x_t$. But in the IMS case this term is significantly more complicated but still feasible using derivations from section 3.5.2 of Lutkepohl (1991). Specifically if we define the $(nm+1) \times (nm+1)$ matrix

[8] If more than one model is being used to construct $f(X_{t+\tau}, \hat{\beta}_t)$ (so that $\hat{\beta}_t = (\hat{\beta}_t^{(1)\prime}, \hat{\beta}_t^{(2)\prime})'$), then B is the block diagonal matrix $diag(B^{(1)}, B^{(2)})$ and hence a consistent estimate is $\hat{B} = diag(\hat{B}^{(1)}, \hat{B}^{(2)})$.

[9] If more than one model is being used to construct $f_{t+\tau}(\hat{\beta}_t)$ (so that $\hat{\beta}_t = (\hat{\beta}_t^{(1)\prime}, \hat{\beta}_t^{(2)\prime})'$), then $h_{t+\tau}$ is the stacked vector $(h_{t+\tau}^{(1)\prime}, h_{t+\tau}^{(2)\prime})'$ and hence the appropriate estimator is $\hat{h}_{t+\tau} = (\hat{h}_{t+\tau}^{(1)\prime}, \hat{h}_{t+\tau}^{(2)\prime})'$.

[10] If $f(X_{t+\tau}, \beta)$ is non-differentiable see McCracken (2004) for an alternative estimator.

$$W = \begin{pmatrix} 1 & 0 & ... & 0 \\ & \Lambda & & \\ 0 & I_{n(m-1)} & & 0 \end{pmatrix},$$

the $n \times (nm+1)$ selection matrix $J_1 = (0_{n\times 1}, I_n, 0_{n\times n(m-1)})$, and $\Phi_i = J_1 W^i J_1'$ we obtain

$$F = -\iota' \sum_{i=0}^{\tau-1} (E x_t' (W')^{\tau-1-i} \otimes \Phi_i).$$

If we use the estimator $\hat{\Lambda}_T$ to obtain $\hat{W}$ and $\hat{\Phi}_i$ and define $\bar{x} = (P-3)^{-1} \sum_{t=R}^{T-4} x_t$, a straightforward estimator of F is simply

$$\hat{F} = -\iota' \sum_{i=0}^{\tau-1} (\bar{x}' (\hat{W}')^{\tau-1-i} \otimes \hat{\Phi}_i).$$

Interestingly, for three special cases estimating Ω is as simple as using the estimate $\hat{\Omega} = \hat{S}_{ff}$. This arises when the second and third terms in equation (12), those due to estimation error, cancel and hence we say the estimation error is asymptotically irrelevant.

Case 1. If $\pi = 0$, then both λ_{fh} and λ_{hh} are zero and hence $\Omega = S_{ff}$. This case arises naturally when the sample split is chosen so that the number of out-of-sample observations is small relative to the number of in-sample observations. Chong and Hendry (1986) first observed that parameter estimation error is irrelevant if P is small relative to R.

Case 2. If $F = 0$, then $\Omega = S_{ff}$. This case arises under certain very specific circumstances but arises most naturally when the measure of "accuracy" is explicitly used when estimating the model parameters and holds regardless of whether DMS or IMS forecasts are used. The canonical example is the use of a quadratic loss function (MSE) to evaluate the accuracy of forecasts from two non-nested models estimated by OLS. In this situation, the F term equals zero and estimation error is asymptotically irrelevant.

Case 3. There are instances where $-S_{fh}B'F' = FBS_{hh}B'F'$ and hence under the recursive scheme, estimation error is asymptotically irrelevant. In this case, it isn't so much than any particular term equals zero but that the sum of the components just happens to cancel to zero. One such example is our test of zero mean prediction error so long as that model contains an intercept and holds regardless of whether DMS or IMS forecasts are used. See West (1996, 2006) and West and McCracken (1998) for other examples.

But more generally, all three components of Ω will have to be estimated and in particular F will have to be estimated. Fortunately, the formula for F associated with a test of zero mean prediction error can be generalized fairly easily in many useful cases. Specifically,

if we assume that the moment function of interest satisfies $f(X_{t+\tau}, \hat{\beta}_t) = \tilde{f}(X_{t+\tau}, y_{1,t+\tau} - \hat{y}_{1,t+\tau}) = \tilde{f}_{t+\tau}(\hat{\varepsilon}_{1,t+\tau})$, and hence the parameter estimate only affects the function through the forecast error (as it does for tests of zero mean prediction error, efficiency, and serial correlation), we obtain

$$F = -\iota' \sum_{i=0}^{\tau-1} (E(\frac{\partial \tilde{f}_{t+\tau}(\varepsilon_{1,t+\tau})}{\partial \varepsilon_1} x_t')(W')^{\tau-1-i} \otimes \Phi_i).$$

For this scenario the same formula for $\hat{F}$ applies if we continue to use the estimator $\hat{\Lambda}_T$ to obtain $\hat{W}$ and $\hat{\Phi}_i$ and redefine $\bar{x}$ as $(P-3)^{-1} \sum_{t=R}^{T-4} \frac{\partial \tilde{f}_{t+\tau}(\hat{\varepsilon}_{1,t+\tau})}{\partial \varepsilon_1} x_t'$.

3.3 Nested Model Comparisons

In the theoretical discussion above the statistic $(P-\tau+1)^{-1/2} \sum_{t=R}^{T-\tau} f(X_{t+\tau}, \hat{\beta}_t)$ was asymptotically normal with variance Ω. To achieve this result, however, it must be the case that Ω is positive – a point emphasized directly in West (1996) through his assumptions. Unfortunately Ω is not trivially positive in certain practical situations including tests of equal forecast accuracy or encompassing when the two nested models are being compared. To see the issue consider a simple example in which DMS forecasts from the autoregressive model

$$y_{1,t} = \beta_0^{(1)} + \beta_1^{(1)} y_{1,t-\tau} + \varepsilon_{1,t}^{(1)} = \beta^{(1)\prime} x_{t-\tau}^{(1)} + \varepsilon_{1,t}^{(1)}$$

are being compared to those from

$$y_{1,t} = \beta_0^{(2)} + \beta_1^{(2)} y_{1,t-\tau} + \beta_2^{(2)} y_{2,t-\tau} + \beta_3^{(2)} y_{3,t-\tau} + \varepsilon_{1,t}^{(2)} = \beta^{(2)\prime} x_{t-\tau}^{(2)} + \varepsilon_{1,t}^{(2)}$$

in terms of their mean squared errors. Specifically suppose that both models are estimated by OLS and we define $(\hat{\varepsilon}_{1,t+\tau}^{(j)})^2 = (y_{1,t+\tau} - \hat{\beta}_t^{(j)\prime} x_{t-\tau}^{(j)})^2$ for the restricted and unrestricted models $j = 1, 2$, respectively. Under the null hypothesis that $E((\varepsilon_{1,t+\tau}^{(1)})^2 - (\varepsilon_{1,t+\tau}^{(2)})^2) = 0$, Clark and McCracken (2001, 2005a) and McCracken (2007) show that the statistic $(P-\tau+1)^{-1/2} \sum_{t=R}^{T-\tau} f(X_{t+\tau}, \hat{\beta}_t)$ with $f(X_{t+\tau}, \hat{\beta}_t) = (\hat{\varepsilon}_{1,t+\tau}^{(1)})^2 - (\hat{\varepsilon}_{1,t+\tau}^{(2)})^2$ converges in probability to zero and is therefore not asymptotically normal is any useful sense.

To get around this problem they extend the results in West (1996) to a framework that allows for nested model comparisons but do so only when the models being compared are OLS estimated linear models and the forecasts are formed using the DMS approach. Specifically, under assumptions close to those considered in West (1996) they extend West's

Lemma 4.3 by showing

$$\begin{aligned}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t) &= \sum_{t=R}^{T-\tau}((\hat{\varepsilon}_{1,t+\tau}^{(1)})^2-(\hat{\varepsilon}_{1,t+\tau}^{(2)})^2)\\ &= 2\sum_{t=R}^{T-\tau} h_{1,t+\tau}^{(2)\prime}(-JB^{(1)}J'+B^{(2)})H^{(2)}(t)\\ &\quad -\sum_{t=R}^{T-\tau} H^{(2)}(t)'(-JB^{(1)}J'+B^{(2)})H^{(2)}(t)+o_p(1)\\ &= O_p(1),\end{aligned}$$

where $B^{(j)} = (Ex_t^{(j)}x_t^{(j)\prime})^{-1}$, $h_{1,t+\tau}^{(2)} = \varepsilon_{1,t+\tau}^{(2)}x_t^{(2)}$, $H^{(2)}(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{1,s+\tau}^{(2)}$, and $J = (I_{\dim(x_t^{(1)})},0)'$.

Building upon that expansion they derive the asymptotic distributions of the MSE-t $= (P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t)/\hat{S}_{ff}^{1/2}$ and MSE-$F = \sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t)/((P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}(\hat{\varepsilon}_{1,t+\tau}^{(2)})^2)$ tests of equal forecast accuracy as well as the ENC-t and ENC-F tests of forecast encompassing, for which $f(X_{t+\tau},\hat{\beta}_t) = \hat{\varepsilon}_{1,t+\tau}^{(1)}(\hat{\varepsilon}_{1,t+\tau}^{(1)}-\hat{\varepsilon}_{1,t+\tau}^{(2)})$. (Note that, in the Monte Carlo section, we also write out these test statistics using just forecast errors.)

Each of the four statistics have asymptotic distributions that can be represented using stochastic integrals of quadratics of Brownian motion and are, in nearly all instances, not asymptotically normal. While useful, the analytics are derived after restricting attention to DMS forecasts since $(\hat{\varepsilon}_{1,t+\tau}^{(1)})^2-(\hat{\varepsilon}_{1,t+\tau}^{(2)})^2$ (and $\hat{\varepsilon}_{1,t+\tau}^{(1)}(\hat{\varepsilon}_{1,t+\tau}^{(1)}-\hat{\varepsilon}_{1,t+\tau}^{(2)})$) are then only quadratic functions of the parameter estimates and the necessary second order Taylor expansion of $\sum_{t=R}^{T-\tau}((\hat{\varepsilon}_{1,t+\tau}^{(1)})^2-(\hat{\varepsilon}_{1,t+\tau}^{(2)})^2)$ (or $\sum_{t=R}^{T-\tau}\hat{\varepsilon}_{1,t+\tau}^{(1)}(\hat{\varepsilon}_{1,t+\tau}^{(1)}-\hat{\varepsilon}_{1,t+\tau}^{(2)})$), as a function of β_1 and β_2, is straightforward. As detailed in such sources as Clark and McCracken (2005a, 2012a,b), critical values from these asymptotic distributions can easily be obtained by bootstrap methods.

In contrast, if the forecasts were constructed using an IMS approach, the second order Taylor expansion of $\sum_{t=R}^{T-\tau}((\hat{\varepsilon}_{1,t+\tau}^{(1)})^2-(\hat{\varepsilon}_{1,t+\tau}^{(2)})^2)$ as a function of β_1 and β_2 would be much more complicated. In particular, if $\tau = 4$ we know that $\sum_{t=R}^{T-\tau}((\hat{\varepsilon}_{1,t+\tau}^{(1)})^2-(\hat{\varepsilon}_{1,t+\tau}^{(2)})^2)$ is an 8th order polynomial in both β_1 and β_2. That is not to say that a similar collection of asymptotics would not be applicable when IMS methods are used to construct forecasts, rather that the results have not been formally derived anywhere in the literature, in large part because to do so would be algebraically tedious and difficult. For that reason, in section 4 we provide Monte Carlo evidence on these tests when the competing model is a VAR that nests a baseline model AR model and all forecasts are generated using an IMS approach. We will compare the tests against bootstrapped critical values.

3.4 Finite Sample Tests of Predictability

The theoretical results in West (1996) and other studies such as Clark and McCracken (2001, 2005a) focus on testing a null hypothesis of the form $H_0 : Ef(X_{t+\tau}, \beta^*) = 0$. In words, this hypothesis is designed to evaluate the predictive ability of a model (or models) if the true parameter values of the model(s) are known. But as a practical matter, forecasting agents never know the true parameter values of a model and must estimate them. Hence, intuitively, one might prefer to evaluate the predictive content of the model accounting for the fact that the model parameters must be estimated and will never be known. Doing so changes the null hypothesis to something akin to $H_0 : Ef(X_{t+\tau}, \hat{\beta}_t) = 0$, where the estimation error associated with the parameter estimates is now captured under the null hypothesis.

While such a hypothesis could be considered for any function $f(X_{t+\tau}, \hat{\beta}_t)$, including that associated with zero mean prediction error, the hypothesis arises most clearly when comparing two ostensibly nested models. In the theory of Clark and McCracken (2001, 2005a) the null of equal population predictive ability maps directly into testing whether the additional predictors in the unrestricted model are associated with coefficients that are zero and hence the hypothesis is $H_0 : Ef(X_{t+\tau}, \beta^*) = E((\varepsilon^{(1)}_{1,t+\tau})^2 - (\varepsilon^{(2)}_{1,t+\tau})^2) = 0$. In contrast, a null of equal finite sample predictive ability between two nested models is akin to the hypothesis $H_0 : Ef(X_{t+\tau}, \hat{\beta}_t) = E((\hat{\varepsilon}^{(1)}_{1,t+\tau})^2 - (\hat{\varepsilon}^{(2)}_{1,t+\tau})^2) = 0$. For this latter hypothesis, the issue isn't so much whether certain coefficients are zero or not, but whether they are non-zero *and* whether they can be estimated sufficiently precisely in a finite sample to be useful for forecasting.

Even though the null hypothesis $H_0 : Ef(X_{t+\tau}, \hat{\beta}_t) = 0$ is reasonably intuitive, the results in West (1996) or Clark and McCracken (2001, 2005a) do not apply under this null and hence their theory cannot be used to infer the asymptotic distribution of $(P - \tau + 1)^{-1/2} \sum_{t=R}^{T-\tau} f(X_{t+\tau}, \hat{\beta}_t)$ under this null regardless of whether DMS or IMS forecasts are used. Instead, theoretical results in Giacomini and White (2006) can be used to conduct inference. Perhaps most importantly from the standpoint of this chapter, their results are applicable regardless of whether DMS or IMS forecasts are being used. Specifically, Giacomini and White (2006) show that so long as all parameter estimates are estimated using a small rolling window of observations (small in the sense that R is finite and P diverges to

infinity), $(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t) \rightarrow^d N(0, S_{\hat{f}\hat{f}})$ under the null hypothesis

$$\lim_{P\rightarrow\infty}(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} Ef(X_{t+\tau},\hat{\beta}_t)=0,$$

where $S_{\hat{f}\hat{f}} = \lim_{P\rightarrow\infty} Var(P^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t))$.

The theoretical results in Giacomini and White (2006) can be used for a wide range of applications, including tests of equal forecast accuracy between nested or non-nested models, zero mean prediction error, forecast encompassing, equal log predictive densities, etc. So long as the statistic takes the form $(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} f(X_{t+\tau},\hat{\beta}_t)$, and all parameters are estimated using a small rolling window of observations, normal critical values can be used to conduct inference on the null hypothesis $\lim_{P\rightarrow\infty}(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau} Ef(X_{t+\tau},\hat{\beta}_t)=0$. Building on Giacomini and White (2006), Amisano and Giacomini (2007) formally develop such tests of equal accuracy of density forecasts, using a logarithmic scoring rule (i.e., the log predictive score). Their tests encompass a t-test applied to the mean score differential and weighted likelihood ratio tests that permit concentration on particular areas of the predictive density.

It's important to also note that, under the asymptotics of Giacomini and White (2006), there are almost no restrictions on how the parameters are estimated. The parameters can be estimated by OLS but can also be estimated using Bayesian methods so long as a small rolling window of R observations is used to estimate the parameters at each forecast origin. Univariate as well as VAR models are permitted.

The only other set of results that apply to tests of the null of finite sample predictive ability are those considered in Clark and McCracken (2011a). There they develop a procedure for testing the null hypothesis $\lim_{P,R\rightarrow\infty}\sum_{t=R}^{T-\tau} E((\hat{\varepsilon}^{(1)}_{1,t+\tau})^2-(\hat{\varepsilon}^{(2)}_{1,t+\tau})^2)=0$ against an alternative in which the unrestricted model 2 is more accurate than the restricted model 1. As was the case for their previous work on tests of equal population predictive ability between nested models, the asymptotic distribution is non-standard and in particular is not asymptotically normal. Instead a bootstrap approach to inference is shown to be asymptotically valid when OLS estimated, linear DMS forecasts are being compared. IMS forecasts from univariate or VAR models are not allowed again due to the fact that the forecast errors are a complicated function of the parameter estimates.

4 Monte Carlo Evidence

In this section we examine the finite sample properties of some of the inference approaches described above. In the interest of brevity and computational tractability, we focus our evaluation on the subset of inference problems that seem to be most common in the VAR forecasting literature. In particular, we examine tests that compare forecasts from VARs and BVARs to baseline, nested AR models, with all forecasts obtained by iteration. We consider both point and density forecasts, generated under both recursive and rolling estimation schemes.

In these experiments, we use trivariate data-generating processes (DGPs) parameterized on the basis of estimates of AR(4) and VAR(4) models in U.S. data on GDP growth, inflation in the price index for personal consumption expenditures (PCE), and the 3-month Treasury bill rate, for an estimation sample of 1961-2007.[11] We consider one set of experiments in which the DGP is a set of AR(4) models for each variable, such that forecasts from the AR models should be at least as accurate as forecasts from a VAR or BVAR. We consider another set of experiments in which the DGP is a VAR(4), such that, with a long enough data sample, forecasts from the VAR or BVAR should be more accurate than forecasts from the AR models. While we would like to be able to include experiments in which the forecasts from the AR and VAR models are equally accurate in the finite sample, there is no easy way of parameterizing a VAR-based DGP to make the forecasts equally accurate in the finite sample. Such a finite-sample evaluation is an important topic for future research. In this paper, we focus on the more tractable DGP settings just described.

In all experiments, we take forecasts from AR(4) models as benchmarks for each variable, and compare against them forecasts from a VAR(4) and a BVAR(4). In the evaluation of point forecasts, to make bootstrap evaluation computationally tractable, we generate the forecasts without simulation. The AR and VAR point forecasts are obtained by using OLS estimates of coefficients and simple iteration of forecasts based on the OLS coefficients. In the case of the BVAR, we also obtain forecasts by iteration, using just the posterior mean coefficient estimates obtained with the analytical solution for the Normal-Wishart posterior. At horizons greater than 1 period, proper Bayesian methods would require simulation of the posterior distribution, but Carriero, Clark, and Marcellino (2012) find that multi-step point

[11] We deliberately end the sample before the financial crisis, to avoid some of the sizable effects the severe recession and slow recovery had on VAR coefficient estimates.

forecasts based on just the posterior mean coefficients are essentially the same as properly simulated point forecasts.

In the evaluation of density forecasts, because the density forecasts are obtained by simulation, computational constraints lead us to avoid bootstrapping and instead limit our analysis to tests compared against asymptotic critical values. We use Bayesian simulation methods to produce the density forecasts, using a extremely loose prior to effectively implement OLS for AR and VAR forecasts and an informative prior to obtain BVAR forecasts.

We proceed by detailing the test statistics and critical values used, our implementation of Bayesian methods, and the data-generating processes and other aspects of experiment design. We then present the results.

4.1 Test statistics and critical values

For the evaluation of point forecasts, we consider four test statistics, consisting of both F-type and t-tests of equal MSE and forecast encompassing. To define these tests for a given variable y_j of the VAR of interest, let model 1 denote the benchmark model for the given variable and model 2 denote the alternative model, $\hat{d}_{t+\tau} = (\hat{\varepsilon}^{(1)}_{j,t+\tau})^2 - (\hat{\varepsilon}^{(2)}_{j,t+\tau})^2$, $\hat{c}_{t+\tau} = \hat{\varepsilon}^{(1)}_{j,t+\tau}(\hat{\varepsilon}^{(1)}_{j,t+\tau} - \hat{\varepsilon}^{(2)}_{j,t+\tau})$, and $\text{MSE}^{(i)} = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}(\hat{\varepsilon}^{(i)}_{j,t+\tau})^2$. Let $\hat{S}_{dd}$ and $\hat{S}_{cc}$ denote the long-run variances of $\hat{d}_{t+\tau}$ and $\hat{c}_{t+\tau}$, computed using a rectangular kernel, $\tau-1$ lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997). For each variable, the four test statistics for point forecasts are computed as:

$$\begin{aligned}
\text{MSE-}F &= \left(\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau}\right)/\text{MSE}^{(2)} \\
\text{MSE-}t &= \left((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau}\right)/\hat{S}_{dd}^{1/2} \\
\text{ENC-}F &= \left(\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}\right)/\text{MSE}^{(2)} \\
\text{ENC-}t &= \left((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}\right)/\hat{S}_{cc}^{1/2}
\end{aligned}$$

(note that the ENC-t test corresponds to the MSE-adjusted test of Clark and West (2006, 2007)). We compare the t-tests against standard normal critical values and all four tests against bootstrapped critical values. We consider two different bootstrap methods, one we refer to as a restricted VAR bootstrap and the other a fixed regressor bootstrap, both

detailed below.

For the evaluation of density forecasts, we consider just one test statistic, a t-test for equal average log predictive score (based on the results of Amisano and Giacomini (2007)), which we will identify as the score-t test. While we focus on this test because the log score is the broadest measure of density accuracy, we certainly recognize that other density-based tests — such as Berkowitz's (2001) test based on normalized forecast errors, Christofferson's (1998) test for interval forecasts, and a test for equality of the mean cumulative ranked probability score of Gneiting and Raftery (2007) — are also of interest. The development and evaluation of such tests for application to VAR forecasts is an important topic for future research.

For variable y_j, letting $\hat{s}_{t+\tau}^{(i)}$ denote the log predictive score of model i (as described below, computed from a simulated forecast distribution), $\hat{d}_{t+\tau} = \hat{s}_{t+\tau}^{(2)} - \hat{s}_{t+\tau}^{(1)}$, and $\hat{S}_{dd} =$ the long-run variance of $\hat{d}_{t+\tau}$ computed using a rectangular kernel, $\tau - 1$ lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997), the test is computed as

$$\text{score-}t = \left((P - \tau + 1)^{-1/2} \sum_{t=R}^{T-\tau} \hat{d}_{t+\tau} \right) / \hat{S}_{dd}^{1/2}.$$

We compare the t-test against standard normal critical values. As noted above, our simulation-based approach to computing density forecasts makes the evaluation of bootstrap-based inference infeasible from a computational perspective.

Based on 2000 Monte Carlo draws, we report the percentage of Monte Carlo trials in which the null of no predictive content is rejected — the percentage of trials in which the sample test statistics exceed the critical values (conducting one-sided tests in all cases). In the reported results, the tests are compared against 10% critical values. Using 5% critical values yields similar findings.

In most cases, our Monte Carlo results apply to tests of equal accuracy in population. For point forecasts, both the restricted VAR and fixed regressor bootstraps generate critical values for tests of the null of equal population predictive ability. Similarly, comparing the ENC-t test against standard normal critical values should be seen as an assessment of predictive content in the population, since, at least in the direct multi-step forecasting case, the normal critical values seem to be fairly close approximations of the quantiles of an asymptotic distribution for the null of no predictive content in population (Clark and West (2007)).

However, comparing the MSE-t test against standard normal critical values may be seen as a test of equal accuracy in the finite sample, since, under the asymptotics of Giacomini and White (2006) and a null of equal accuracy in the finite sample, the t-test has a standard normal distribution. The same applies to the score-t test used to test the equality of average log predictive scores. That said, as noted above, the asymptotic normality results of Giacomini and White (2006) and Amisano and Giacomini (2007) require that forecasts be generated with a rolling sample approach to estimation. We will nonetheless evaluate the efficacy of MSE-t and score-t tests compared against normal critical values for both recursive and rolling estimation schemes. We leave as a subject for future research the further investigation of inference of equal accuracy in the finite sample. As we indicated above, this investigation will require finding a way to simulate data to ensure the model-based forecasts are equally accurate in the finite sample. It may also involve developing other tests or bootstrap approaches for testing equal accuracy in the finite sample.

4.1.1 Restricted VAR bootstrap

One bootstrap method we consider is patterned on the approach first used in Kilian (1999) and applied in the univariate forecast evaluation of such studies as Clark and McCracken (2001, 2005a). With the null forecasting model for each variable taken to be an AR model, the bootstrap model is a set of AR(4) equations for each variable, with coefficients estimated by OLS using the full sample of data. Bootstrapped time series on the vector y_t are generated by drawing with replacement from the vector of OLS residuals and using the autoregressive structures of the AR models to iteratively construct data.[12] In each of 499 bootstrap replications, the bootstrapped data are used to recursively estimate AR, VAR, and BVAR models and generate forecasts. In each replication, the resulting forecasts are then used to calculate forecast test statistics. Critical values are simply computed as percentiles of the bootstrapped test statistics.

4.1.2 Fixed regressor bootstrap

The other bootstrap method we consider is patterned on the fixed regressor wild bootstrap developed in Goncalves and Kilian (2004) and extended to out-of-sample tests by Clark and McCracken (2012a). We generate artificial data on the predictands using (a) fitted values from AR(4) models for each variable, which correspond to the restricted forecasting

[12]The initial observations are selected by sampling from the actual data as in Stine (1987).

models, and (b) residuals from the relevant unrestricted forecasting model, which is either the VAR or BVAR. For the data simulation, all of the models used are estimated with the full sample of data. We construct artificial time series y_t^* by adding the vector of fitted values from the AR models to a vector of innovations equal to the innovations from the relevant unrestricted model times a draw from the standard normal distribution.

More specifically, for each variable $y_{i,t}$, we use the full sample of data to estimate an AR(4) model to obtain a vector of coefficients $\hat{\beta}_i$. We also estimate the VAR and BVAR models, to obtain vectors of residuals $\hat{\varepsilon}_t^{(\text{VAR})}$ and $\hat{\varepsilon}_t^{(\text{BVAR})}$. For comparing VAR forecasts against the AR benchmark, we repeatedly construct a vector of simulated time series $y_{i,t}^*$ using

$$y_{i,t}^* = \hat{\beta}_{i,0} + \sum_{j=1}^{4} \hat{\beta}_{i,j} y_{i,t-j} + \eta_t \hat{\varepsilon}_{i,t}^{(\text{VAR})},$$

where η_t is a draw from a N(0,1) distribution, and the same draw is used for each variable of the VAR. To obtain a bootstrapped distribution of test statistics, for each data draw we obtain pseudo-AR model forecasts by regressing $y_{i,t}^*$ on lags of $y_{i,t}$, using lagged values of $y_{i,t}$ to iteratively construct forecasts of horizons 1 through τ, and forming forecast errors as $y_{i,t+\tau}^*$ less the forecast. We obtain pseudo-VAR forecasts by regressing each variable $y_{i,t}^*$ on lags of the vector y_t, using lagged values of y_t to iteratively construct forecasts of horizons 1 through τ, and forming forecast errors for each variable as $y_{i,t+\tau}^*$ less the forecast. For comparing BVAR forecasts against the AR benchmark, the procedure is the same, except that residuals from the BVAR are used in generating artificial data $y_{i,t}^*$, and simulated forecasts are computed for the BVAR rather than the VAR. For each draw of simulated forecast errors, we compute the four test statistics of interest, to obtain a bootstrapped distribution against which we compare the sample test statistics from the Monte Carlo-generated data set.

4.2 Details on implementation of Bayesian methods

The BVAR takes the form described in section 2.2.2:

$$\begin{aligned} Y_t &= C + A(L)Y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \Sigma) \\ &= \Lambda x_{t-1} + \varepsilon_t. \end{aligned}$$

For this model, we use the Normal-Wishart prior and posterior detailed in such studies as Kadiyala and Karlsson (1997) and Banbura, Giannone, and Reichlin (2010).

The prior takes a conventional Minnesota form, without cross-variable shrinkage, in which the prior for the VAR coefficients is normally distributed:

$$\text{vec}(\Lambda) \sim N(\text{vec}(\underline{\mu}_\Lambda), \underline{\Omega}_\Lambda).$$

Letting $\Lambda_l^{(ij)}$ denote the coefficient on lag l of variable j in equation i and $\Lambda_l^{(ij)}$ with l=0 denote the intercept of equation i, we define the prior mean and variance as follows:

$$\underline{\mu}_\Lambda \text{ such that } E[\Lambda_l^{(ij)}] = 0 \; \forall \; i, j, l \tag{13}$$

$$\underline{\Omega}_\Lambda \text{ such that } V[\Lambda_l^{(ij)}] = \begin{cases} \frac{\theta^2}{l^2}\frac{\sigma_i^2}{\sigma_j^2} & \text{for } l > 0 \\ \varepsilon^2 \sigma_i^2 & \text{for } l = 0 \end{cases}. \tag{14}$$

In our BVAR implementation, we set the tightness hyperparameter θ at 0.5 and the intercept hyperparameter ε to 1000, and we set the scale parameters σ_i^2 at estimates of residual variances from AR(4) models from the estimation sample available at each forecast origin. In using the same simulation code for producing density forecasts from AR and VAR models, we make the prior uninformative by setting $\theta = 1000$ and dropping the l^2 in the denominator of the prior variance.

For the evaluation of point forecasts from the BVAR, we obtain forecasts using the posterior mean value of the coefficients and the iterative approach to forecasting at horizons greater than 1 period. For the evaluation of density forecasts, we simulate 2000 draws of forecasts from the posterior distribution as follows: for each draw of coefficients and Σ from their posterior distributions, we (1) draw shocks to y_t over the forecast horizon with variance-covariance matrix equal to the draw of Σ, and (2) iterate forward using the autoregressive structure of the VAR and the draw of the VAR coefficients to obtain artificial $y_{t+\tau}$. At each forecast horizon, for each variable $y_{i,t+\tau}$, we compute the log score based on the actual outcome for $y_{i,t+\tau}$ and an empirical estimate of the forecast density obtained with a Gaussian kernel and the draws of forecasts from the posterior distribution.

4.3 Monte Carlo Design

For two different trivariate DGPs — one in which the true model consists of AR(4) processes for each variable and the other in which the true model is a VAR(4) — we generate data using independent draws of innovations from the normal distribution and the autoregressive structure of the DGP. The initial observations necessitated by the lag structure of each DGP are generated with draws from the unconditional normal distribution implied by the DGP.

With quarterly data in mind, we report results for forecast horizons of 1, 2, 4, and 8 periods ahead, with sample sizes of R, P = 80,40; 80,80; and 80,120. We obtained qualitatively similar results (available upon request) for sample sizes of R, P = 40,80; 40,120; 120,40; and 120,80.[13]

The parameters of the DGPs are set on the basis of 1961-2007 model estimates for U.S. data on GDP growth, PCE inflation, and the 3-month Treasury bill rate. In DGP 1, the true model consists of AR(4) processes for each variable, with coefficients given in Table 1 and an error variance-covariance matrix of:

$$\text{var}(\varepsilon_t) = \begin{pmatrix} 9.81 & & \\ -0.05 & 1.35 & \\ 0.32 & 0.16 & 0.48 \end{pmatrix}.$$

In DGP 2, the true model is a VAR(4), with coefficients given in Table 2 and an error variance-covariance matrix of:

$$\text{var}(\varepsilon_t) = \begin{pmatrix} 8.35 & & \\ -0.06 & 1.16 & \\ 0.36 & 0.13 & 0.41 \end{pmatrix}. \tag{15}$$

4.4 Results

Tables 3 and 4 provide empirical rejection rates from Monte Carlo experiments with DGP 1 based on a recursive estimation scheme, and Tables 5 and 6 provide rejection rates from DGP 1 experiments using a rolling estimation scheme. Tables 7 and 8 provide recursive scheme results from experiments with DGP 2. Each table provides results for three different sample size combinations, for either VAR versus AR forecasts or BVAR versus AR forecasts. Each table covers all three variables being forecast and forecast horizons of 1, 2, 4, and 8 periods ahead. Because the BVAR and VAR results are broadly similar, in the ensuing discussion we focus on the VAR versus AR comparisons, in the interest of brevity.[14] The test statistic and source of critical values are indicated in the first column of each table.

In the DGP 1 experiments, in which the true model is a set of AR equations, and therefore the null hypothesis of no predictive accuracy in population is satisfied, the bootstrap-based tests yield rejection rates reasonably close to the nominal size of 10 percent. However, the size results seem to be more consistently close to 10 percent when critical values are

[13] In results for these other sample sizes, the most notable difference is that, with DGP experiments using R, P = 40,80, the tests sometimes display slightly larger size distortions.

[14] The largest difference seems to be that, in the DGP 1 results, the ENC-t test compared against standard normal critical values yields a higher rejection rate when applied to BVAR versus AR forecasts than when applied to VAR versus AR forecasts.

obtained from the VAR bootstrap (VARBS) than when critical values are obtained from the fixed regressor bootstrap (FRBS). While the FRBS rejection rates are typically close to 10 percent, there are a number of instances in which the rejection rates are somewhat above or below 10 percent. Consider, for example, the VAR versus AR results in Table 3 for a sample size combination of R, P = 80,80. For most variables and forecast horizons, the empirical size of the MSE-F test based on the FRBS is 10 or 11 percent. However, with variable 1, the FRBS-based size is about 13 percent at horizons of 4 and 8, while with variable 3, the FRBS-based size is about 7 percent at horizons of 4 and 8. Using critical values from the VARBS more consistently yields empirical rejection rates close to 10 percent. In the same example, across the three variables and four horizons, the size of the MSE-F test based on the VARBS ranges from 9.2 to 12.6 percent. Similarly, as indicated in Table 3, for many variables and horizons, the empirical size of the ENC-F test based on the FRBS is 10 or 11 percent. However, with variable 2 and a 2-step-ahead horizon, the FRBS-based size spikes to almost 16 percent. Again, size performance is more even with the VARBS, yielding ENC-F rejection rates that range from 8.1 to 12.1 percent (compared to a range of 9.2 to 15.9 percent with the FRBS).

Given the bootstrap method, size results are broadly similar for the MSE-F, MSE-t, ENC-F, and ENC-t tests. The most noticeable difference is that, when the fixed regressor bootstrap is used to generate critical values, the t-tests are less prone to oversizing and more prone to undersizing than their F-test counterparts. Consider, for example, the FRBS-based encompassing test results in the Table 3 experiments with R, P = 80,80. The size of the ENC-F test ranges from 9.2 to 15.9 percent; the size of the ENC-t test varies from 5.2 to 10.9 percent. In corresponding results in the experiments with R, P = 80,40, the empirical rejection rate of the ENC-F test ranges from 9.6 to 15.8 percent; the size of the ENC-t test varies from 6.5 to 11.2 percent.

When the t-tests for point forecasts are compared against standard normal critical values, the MSE-t and ENC-t tests behave fairly differently, in line with expectations. As noted above, the MSE-t test compared against normal critical values is in effect (subject to the conditions of Giacomini and White (2006)) a test of the null of equal accuracy in the finite sample, rather than in population. Accordingly, when the AR model is the true model and therefore likely to be more accurate than a VAR in the finite sample, the (one-sided) MSE-t test compared against standard normal critical values usually yields a rejection rate

well below 10%. How much below depends on the forecast horizon: the rejection rate tends to rise with the forecast horizon, particularly with smaller P, probably due to challenges in estimating precisely the HAC standard deviation in the denominator of the test statistic. For example, as shown in Table 3's results with R, P = 80,120, the rejection rate of the MSE-t test ranges from 0.0 to 3.8 percent. In results with R, P = 80,40, the rejection rate varies from 1.3 to 8.8 percent.

In contrast, the ENC-t test compared against normal critical values is approximately a test of the null of no predictive content in population and may therefore be expected to yield rejection rates in the DGP 1 experiments that are close to the nominal size of 10 percent. The test is indeed about correctly sized when P is relatively large, but for smaller P, it tends to be oversized at longer forecast horizons. For instance, in Table 3's results with R, P = 80,120, the rejection rate of the ENC-t test based on standard normal critical values ranges from 8.4 to 11.1 percent. In results with R, P = 80,40, the rejection rate varies from 10.5 to 15.4 percent.

Turning to the density forecast results, comparing the score-t test against standard normal critical values yields rejection rates broadly similar to those obtained for the MSE-t test for point forecasts. Like the MSE-t test, the score-t test compared against normal critical values is in effect (subject to the conditions of Amisano and Giacomini (2007)) a test of the null of equal (density) accuracy in the finite sample, rather than in population. Accordingly, when the AR model is the true model and therefore likely to be more accurate than a VAR in the finite sample, the (one-sided) score-t test compared against standard normal critical values usually yields a rejection rate well below 10%, more so at shorter horizons than longer horizons. As an example, in Table 3's results with R, P = 80,120, the rejection rate of the score-t test ranges from 0.1 to 3.2 percent. Taken at face value, these findings suggest that tests for equal accuracy in density forecasts are likely to yield results similar to tests for equal accuracy of point forecasts. In practice, however, results could differ due to complications outside the scope of our experiments: non-Gaussian errors, time-varying volatility (particularly if one model captures it and the other does not), or model misspecification.

The results in Tables 5 and 6 indicate that, with DGP 1, using a rolling scheme for estimation and forecasting yields results qualitatively similar to those obtained using a recursive scheme. It is again the case that the bootstrap-based tests yield rejection rates

reasonably close to the nominal size of 10 percent, with the VAR bootstrap being more consistent across specifications than the fixed regressor bootstrap. As an example, in Table 5's results for the VAR versus the AR with R, P = 80,40, the size of the MSE-F test ranges from 9.3 to 15.8 percent under the FRBS and from 9.7 to 12.0 percent under the VARBS. In addition, size results are similar across the different tests, with the biggest difference being that, when critical values are obtained from the FRBS, the t-tests are less prone to oversizing and more prone to undersizing than their F-test counterparts. In the same example, the size of the MSE-t test compared against standard normal critical values ranges from 8.0 to 14.6 percent. Finally, under the rolling scheme, it is also the case that, when the t-tests are compared against standard normal critical values, the MSE-t test usually yields a rejection rate well below 10 percent, while the ENC-t test has a rejection rate ranging from about 9 to 16 percent. As in the recursive case, the score-t test for density forecasts yields a rejection rate very similar to the normal critical values-based rejection rate of the MSE-t test for equal accuracy of point forecasts.

Returning to results based on the recursive estimation and forecasting scheme, in the DGP 2 experiments of Tables 7 and 8, in which the true model is a VAR, the bootstrap-based tests yield reasonably high rejection rates, particularly at shorter forecast horizons. Broadly, rejection rates are similar for the VAR-based and fixed regressor bootstraps. It is also the case that rejection rates are broadly similar for tests of VAR accuracy versus the AR and for tests of BVAR accuracy versus the AR. Accordingly, in the interest of brevity, in the following discussion we focus on test results for VAR forecasts obtained with the VAR-based bootstrap and with standard normal critical values.

Consider, for example, results in Table 7 for variable 2 and R, P = 80,80 obtained with the VAR bootstrap. At the 1-step horizon, the rejection rates are 82.6 percent for MSE-F, 84.1 percent for MSE-t, 94.4 percent for ENC-F, and 92.8 for ENC-t. At the 4-step horizon, the corresponding rejection rates are, respectively, 57.3, 53.1, 68.8, and 64.2 percent. As might be expected, power systematically rises as the size of the forecast sample (P) rises and falls as the forecast horizon falls. For instance, using the VARBS, the rejection rate for the MSE-F test applied to 1-step ahead forecasts of variable 2 rises from 63.6 percent for $P = 40$ to 82.6 percent for $P = 80$ to 92.8 percent for $P = 120$. Using the VARBS, the rejection rate for the MSE-F test applied to forecasts of variable 2 falls from 82.6 percent for $h = 1$ to, respectively, 68.9, 57.3, and 30.1 percent for horizons 2, 4, and 8.

For the different tests compared to bootstrap critical values, their power rankings broadly line up with the results of Clark and McCracken (2001, 2005a,b) on univariate forecasts. The ENC-F test rejects the null of equal accuracy in population more frequently than does the MSE-F test. Consider the VAR bootstrap-based results for forecasts of the second variable in experiments with R, P = 80,80 (Table 7). The ENC-F rejection rate is 94.4 percent at the 1-step horizon and 68.8 percent at the 4-step horizon; the corresponding MSE-F rejection rates are 82.6 and 57.3 percent, respectively. It is also the case that the ENC-t test rejects the null more frequently than does the MSE-t test. In the same example, the ENC-t rejection rate is 92.8 and 64.2 percent at horizons of 1 and 4 periods, respectively, compared to MSE-t rejection rates of 84.1 and 53.1 percent. As these examples suggest, it is also usually the case that power is at least as great with the F-type tests as with t-type tests, with larger differences at longer horizons than shorter horizons.

In the DGP 2 experiments, as in the DGP 1 experiments, the MSE-t and ENC-t tests compared against standard normal critical values behave fairly differently. The ENC-t test compared against normal critical values rejects the null of equal accuracy in population at least as often as does the same test compared against bootstrapped critical values. Consider again forecasts of the second variable in experiments with R, P = 80,80 (Table 7). Comparing the ENC-t test against VARBS critical values yields a rejection rate of 92.8 and 64.2 percent at horizons of 1 and 4 periods, respectively, while comparing the test against standard normal critical values yields corresponding rejection rates of 92.8 and 66.6 percent, respectively. Because the MSE-t test compared against normal critical values is effectively a test of the null of equal accuracy in the finite sample, rather than in population, it systematically rejects less often than do any of the other tests, all of which are more appropriately thought of as tests of equal accuracy in population. Continuing with the same example, the rejection rate for the MSE-t test compared against standard normal critical values is 29.4 percent at the 1-step horizon and 20.2 percent at the 4-step horizon.

Another parallel to the DGP 1 results is that, in the DGP 2 experiments, rejection rates for the score-t test applied to density forecasts are similar to those for the MSE-t test applied to point forecasts and compared to standard normal critical values. For instance, in forecasts of the second variable in experiments with R, P = 80,80, the rejection rate of the score-t test is 28.5 percent at the 1-step horizon and 22.7 percent at the 4-step horizon. The only notable differences in the rejection rates of the two tests occur with variable 1

and longer forecast horizons, for which the score-t test has a rejection rate as much as 10 percentage points higher than the MSE-t test.

To summarize the implications of the Monte Carlo analysis, based on our results for testing VAR forecasts versus AR forecasts, what we refer to as the VAR-based bootstrap (in which the null of no predictive content is imposed on the bootstrap DGP) appears to be reasonably reliable for testing equal accuracy in population. It is harder to know much about the reliability of tests for equal accuracy in the finite sample, based on standard normal critical values, because there is no simple way of parameterizing a Monte Carlo DGP to make AR and VAR forecasts equally accurate in the finite sample. Our results for such tests, based on DGPs in which either the AR or VAR is the model generating the data, may be seen as weakly suggestive that the finite-sample tests will behave as desired. We draw that inference based on our finding that, when the data are generated by AR models, such that there is no predictability in population, rejection rates for the finite sample tests are low.

5 Conclusion

This chapter reviews recent developments in the evaluation of point and density forecasts in the context of vector autoregressions. In particular we focus on testing issues specific to iterated multi-step forecasts made from VARs as compared to direct multi-step forecasts generated by single equation models. We discussed these issues as they relate to the evaluation of forecasts in population (based on true, unknown model coefficients) as well as the evaluation of forecasts in the finite sample (based on estimated model coefficients).

For evaluation in finite-samples, the results of Giacomini and White (2006) broadly apply with VAR forecasts, subject to the requirement that the forecasts be generated with a fixed or rolling window estimation approach. In a VAR context, it would be very difficult to define a data generating-process that satisfied the null hypothesis and thereby permitted an analysis of the performance of the tests in practice. However, we provide Monte Carlo analysis to show that tests of equal accuracy in the finite sample behave as expected when applied under a DGP designed for the null of equal accuracy in population. In particular, the finite-sample tests are much more conservative (in rejection rates) than the tests of equal accuracy in population.

For evaluation in population, the distinction between iterated multi-step and direct

multi-step approaches becomes important. For direct multi-step forecasts, existing results from sources such as West (1996), Clark and McCracken (2001, 2005a) and McCracken (2007) go through directly and simply. For iterated multi-step forecasts and evaluations that do not involve nested models, the results of West (1996) still apply, but the correction to standard errors necessitated by parameter estimation becomes more complicated. For iterated multi-step forecasts and evaluations of nested models, the asymptotic results of Clark and McCracken (2001a, 2005) and McCracken (2007) no longer apply, because the multi-step forecasts are not linear functions of least-squares estimated parameters.

Still, even in this case, it may that the bootstrap methods considered in Clark and McCracken (2001a, 2005) turn out to be reasonably accurate when applied to nested model forecasts from VAR models. Our Monte Carlo experiments bear this out. Overall, our Monte Carlo analysis suggests that bootstrap-based tests of equal accuracy have reasonable properties under the null of equal accuracy in population.

References

Amisano, Gianni, and Rafaella Giacomini (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business and Economic Statistics* 25, 177-190.

Andersson, Michael K., Gustav Karlsson, and Josef Svensson (2007), "The Riksbank's Forecasting Performance," Sveriges Riksbank working paper #218.

Banbura, Michele, Domenico Giannone, and Lucrezia Reichlin (2010), "Large Bayesian Vector Autoregressions," *Journal of Applied Econometrics* 25, 71-92.

Baumeister, Christiane, and Lutz Kilian (2012a), "Real-Time Forecasts of the Real Price of Oil," *Journal of Business and Economic Statistics* 30, 326-336.

Baumeister, Christiane, and Lutz Kilian (2012b), "Real-Time Analysis of Oil Price Risks using Forecast Scenarios," mimeo, University of Michigan.

Baumeister, Christiane, and Lutz Kilian (2012c), "What Central Bankers Need to Know about Forecasting Oil Prices," mimeo, University of Michigan.

Beauchemin, Kenneth R. and Saeed Zaman (2011), "A Medium Scale Forecasting Model for Monetary Policy," Federal Reserve Bank of Cleveland, working paper no. 11-28.

Berkowitz, Jeremy (2001), "Testing Density Forecasts, With Applications to Risk Management," *Journal of Business and Economic Statistics* 19, 465-474.

Bjornland, Hilde C., Karsten Gerdrup, Anne Sofie Jore, Christie Smith, and Leif Anders Thorsrud (2012), "Does Forecast Combination Improve Norges Bank Inflation Forecasts?" *Oxford Bulletin of Economics and Statistics* 74, 163-179.

Carriero Andrea, Todd E. Clark, and Massimiliano Marcellino (2012), "Bayesian VARs: Specification Choices and Forecast Accuracy," *Journal of Applied Econometrics*, forthcoming.

Chong, Yock Y., and David F. Hendry (1986), "Econometric Evaluation of Linear Macroeconomic Models," *Review of Economic Studies* 53, 671-690.

Christoffersen, Peter F. (1998), "Evaluating Interval Forecasts," *International Economic Review* 39, 841-862.

Clark, Todd E. (2011), "Real-Time Density Forecasts from BVARs with Stochastic Volatility," *Journal of Business and Economic Statistics* 29, 327-341.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

Clark, Todd E., and Michael W. McCracken (2005a), "Evaluating Direct Multistep Forecasts," *Econometric Reviews* 24, 369-404.

Clark, Todd E., and Michael W. McCracken (2005b), "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics* 124, 1-31.

Clark, Todd E. and Michael W. McCracken (2008), "Forecasting with Small Macroeconomic VARs in the Presence of Instabilities," in Forecasting in the Presence of Structural Breaks and Model Uncertainty, ed. D.E. Rapach and M.E. Wohar, (Amsterdam:

Elsevier).

Clark, Todd E., and Michael W. McCracken (2009), "Tests of Equal Predictive Ability with Real-Time Data," *Journal of Business and Economic Statistics* 27, 441-454.

Clark, Todd E., and Michael W. McCracken (2010), ""Averaging Forecasts from VARs with Uncertain Instabilities," *Journal of Applied Econometrics* 25, 5-29.

Clark, Todd E., and Michael W. McCracken (2011a), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," manuscript, Federal Reserve Bank of St. Louis.

Clark, Todd E., and Michael W. McCracken (2011b), "Testing for Unconditional Predictive Ability," in *Oxford Handbook of Economic Forecasting*, Michael P. Clements and David F. Hendry, eds., Oxford: Oxford University Press.

Clark, Todd E., and Michael W. McCracken (2012a), "Reality Checks and Comparisons of Nested Predictive Models," *Journal of Business and Economic Statistics* 30, 53-66.

Clark, Todd E., and Michael W. McCracken (2012b), "Advances in Forecast Evaluation," Forthcoming in Handbook of Economic Forecasting, vol. 2.

Clark, Todd E., and Kenneth D. West (2006), "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics* 135, 155-186.

Clark, Todd E., and Kenneth D. West (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

Coletti, Don, and Stephen Murchison (2002), "Models in policy-making," *Bank of Canada Review*, 19-26.

Corradi, Valentina, Norman R. Swanson, and Claudia Olivetti (2001), "Predictive Ability with Cointegrated Variables," *Journal of Econometrics* 105, 315-358.

Doan, Thomas, Robert B. Litterman, and Christopher A. Sims (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews* 3, 1-100.

Geweke, John, and Charles Whiteman (2006), "Bayesian Forecasting," in *Handbook of Economic Forecasting*, Elliott, G., Granger, C., and Timmermann, A. (eds.). North-Holland: Amsterdam.

Giacomini, Rafaella, and Halbert White (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-1578.

Giannone, Domenico, Michele Lenza, and Giorgio Primiceri (2012), "Prior Selection for Vector Autoregressions," mimeo, Free University of Brussels.

Gneiting, Tilmann, and Adrian E. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association* 102, 359-378.

Goncalves, Silvia, and Lutz (2004), "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form," *Journal of Econometrics*, 123, 89-120.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting* 13, 281-91.

Inoue, Atsushi, and Lutz Kilian (2004), "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?" *Econometric Reviews* 23, 371-402.

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

Kapetanios, George, Vincent Labhard, and Simon Price (2008), "Forecast Combination and the Bank of England's Suite of Statistical Foresating Models," *Economic Modelling* 25, 772-792.

Karlsson, Sune (2012), "Forecasting with Bayesian VAR models," forthcoming, *Handbook of Economic Forecasting*, vol. 2, Elsevier.

Kilian, Lutz (1999), "Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?" *Journal of Applied Econometrics* 14, 491-510.

Komunjer, Ivana, and Michael T. Owyang (2012), "Multivariate Forecast Evaluation and Rationality Testing," *Review of Economics and Statistics* 94, 1066-1080.

Koop, Gary (2012), "Forecasting with Medium and Large Bayesian VARs," *Journal of Applied Econometrics*, forthcoming.

Litterman, Robert B. (1986), "Forecasting with Bayesian Vector Autoregressions-Five Years of Experience," *Journal of Business and Economic Statistics* 4, 25-38.

Lutkepohl, Helmut (1991), *Introduction to Multiple Time Series Analysis*, Springer-Verlag, Berlin.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2006), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics* 127, 499-526.

McCracken, Michael W. (2000), "Robust Out-of-Sample Inference," *Journal of Econometrics* 99, 195-223.

McCracken, Michael W. (2004), "Parameter Estimation Error and Tests of Equal Forecast Accuracy Between Non-nested Models," *International Journal of Forecasting* 20, 503-514.

McCracken, Michael W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics* 140, 719-752.

Newey, Whitney K., and Kenneth D. West (1987), "A simple, positive semi–definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703–708.

Schorfheide, Frank (2005), "VAR forecasting under misspecification," *Journal of Econometrics* 128, 99-136.

Sims, Christopher A. (1980), "Macroeconomics and Reality," *Econometrica* 48, 1-48.

Stine, Robert A. (1987), "Estimating Properties of Autoregressive Forecasts," *Journal of*

the American Statistical Association 82, 1072-1078.

West, Kenneth D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-1084.

West, Kenneth D. (2006), "Forecast Evaluation," in *Handbook of Economic Forecasting*, Elliott G., Granger C.W.J., Timmermann, A. (eds), North Holland.

West, Kenneth D., and Michael W. McCracken (1998), "Regression-based Tests of Predictive Ability," *International Economic Review* 39, 817-840.

Wright, Jonathan H. (2012), "Evaluating Real-Time VAR Forecasts with an Informative democratic Prior," forthcoming, *Journal of Applied Econometrics*.

Zellner, Arnold (1971), An Introduction to Bayesian Inference in Econometrics, J. Wiley and Sons, Inc., New York.

Table 1. Monte Carlo DGP 1 coefficients

explanatory variable	$y_{1,t}$ equation	$y_{2,t}$ equation	$y_{3,t}$ equation
$y_{1,t-1}$	0.24	0.00	0.00
$y_{1,t-2}$	0.16	0.00	0.00
$y_{1,t-3}$	-0.05	0.00	0.00
$y_{1,t-4}$	0.03	0.00	0.00
$y_{2,t-1}$	0.00	0.61	0.00
$y_{2,t-2}$	0.00	0.13	0.00
$y_{2,t-3}$	0.00	0.30	0.00
$y_{2,t-4}$	0.00	-0.14	0.00
$y_{3,t-1}$	0.00	0.00	1.30
$y_{3,t-2}$	0.00	0.00	-0.64
$y_{3,t-3}$	0.00	0.00	0.59
$y_{3,t-4}$	0.00	0.00	-0.30
intercept	2.08	0.34	0.30

Notes:
1. The table provides the coefficients of Monte Carlo DGP 1, in the form of a VAR with 0 restrictions to make the model correspond to an AR.
2. The variance-covariance matrix of innovations and other aspects of the Monte Carlo design are described in section 4.3.

Table 2. Monte Carlo DGP 2 coefficients

explanatory variable	$y_{1,t}$ equation	$y_{2,t}$ equation	$y_{3,t}$ equation
$y_{1,t-1}$	0.18	0.04	0.05
$y_{1,t-2}$	0.19	-0.01	0.04
$y_{1,t-3}$	-0.04	-0.03	-0.01
$y_{1,t-4}$	0.05	0.12	0.01
$y_{2,t-1}$	-0.15	0.59	0.06
$y_{2,t-2}$	0.09	0.17	0.10
$y_{2,t-3}$	-0.19	0.28	-0.01
$y_{2,t-4}$	0.14	-0.06	-0.03
$y_{3,t-1}$	0.36	0.17	1.15
$y_{3,t-2}$	-1.73	-0.20	-0.63
$y_{3,t-3}$	1.24	-0.06	0.65
$y_{3,t-4}$	-0.06	0.07	-0.26
intercept	3.47	-0.13	-0.23

Notes:
1. The table provides the coefficients of Monte Carlo DGP 2, in the form of a VAR.
2. The variance-covariance matrix of innovations and other aspects of the Monte Carlo design are described in section 4.3.

Table 3. Monte Carlo Rejection Rates, 10% critical values: DGP 1, Recursive Scheme, VAR vs. AR

	forecasts of $y_{1,t}$				forecasts of $y_{2,t}$				forecasts of $y_{3,t}$			
test, critical values	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$
R,P = 80,40												
MSE-F, VARBS	0.107	0.115	0.107	0.130	0.100	0.106	0.097	0.098	0.105	0.110	0.111	0.108
MSE-F, FRBS	0.113	0.142	0.149	0.148	0.112	0.139	0.118	0.095	0.126	0.120	0.104	0.085
MSE-t, VARBS	0.110	0.114	0.113	0.111	0.105	0.105	0.101	0.100	0.104	0.105	0.110	0.096
MSE-t, FRBS	0.114	0.136	0.124	0.139	0.111	0.115	0.081	0.086	0.123	0.095	0.060	0.071
ENC-F, VARBS	0.096	0.105	0.109	0.134	0.088	0.091	0.091	0.098	0.106	0.111	0.115	0.120
ENC-F, FRBS	0.113	0.132	0.133	0.111	0.113	0.158	0.132	0.104	0.117	0.155	0.134	0.096
ENC-t, VARBS	0.107	0.104	0.104	0.116	0.096	0.096	0.095	0.096	0.105	0.104	0.103	0.104
ENC-t, FRBS	0.112	0.107	0.091	0.106	0.101	0.093	0.073	0.084	0.108	0.079	0.065	0.085
MSE-t, normal	0.013	0.020	0.051	0.088	0.015	0.029	0.050	0.077	0.013	0.021	0.043	0.072
ENC-t, normal	0.110	0.121	0.137	0.143	0.105	0.118	0.128	0.146	0.113	0.124	0.142	0.154
score-t, normal	0.015	0.026	0.046	0.063	0.011	0.030	0.052	0.084	0.012	0.024	0.048	0.087
R,P = 80,80												
MSE-F, VARBS	0.126	0.110	0.106	0.103	0.116	0.101	0.108	0.115	0.092	0.093	0.101	0.104
MSE-F, FRBS	0.101	0.116	0.134	0.132	0.106	0.117	0.107	0.101	0.092	0.083	0.074	0.072
MSE-t, VARBS	0.118	0.108	0.105	0.100	0.113	0.101	0.108	0.113	0.091	0.095	0.102	0.097
MSE-t, FRBS	0.103	0.121	0.122	0.121	0.109	0.117	0.076	0.073	0.099	0.073	0.043	0.053
ENC-F, VARBS	0.084	0.081	0.086	0.098	0.096	0.088	0.104	0.107	0.088	0.097	0.102	0.121
ENC-F, FRBS	0.105	0.122	0.115	0.092	0.120	0.159	0.140	0.104	0.098	0.138	0.115	0.093
ENC-t, VARBS	0.090	0.098	0.088	0.101	0.104	0.096	0.104	0.109	0.086	0.098	0.098	0.108
ENC-t, FRBS	0.090	0.095	0.075	0.087	0.109	0.084	0.061	0.071	0.088	0.061	0.052	0.067
MSE-t, normal	0.002	0.005	0.019	0.044	0.003	0.005	0.022	0.052	0.004	0.005	0.022	0.046
ENC-t, normal	0.085	0.094	0.092	0.121	0.098	0.098	0.111	0.127	0.081	0.095	0.106	0.126
score-t, normal	0.005	0.010	0.025	0.042	0.003	0.011	0.025	0.045	0.003	0.007	0.015	0.047
R,P = 80,120												
MSE-F, VARBS	0.131	0.122	0.113	0.112	0.121	0.113	0.117	0.119	0.108	0.110	0.105	0.096
MSE-F, FRBS	0.093	0.110	0.135	0.147	0.098	0.105	0.108	0.091	0.104	0.082	0.063	0.060
MSE-t, VARBS	0.119	0.117	0.116	0.104	0.113	0.111	0.119	0.111	0.110	0.110	0.105	0.092
MSE-t, FRBS	0.096	0.127	0.129	0.133	0.103	0.123	0.077	0.066	0.111	0.084	0.041	0.038
ENC-F, VARBS	0.085	0.097	0.106	0.115	0.099	0.099	0.106	0.113	0.101	0.106	0.111	0.108
ENC-F, FRBS	0.106	0.131	0.135	0.106	0.115	0.149	0.138	0.098	0.119	0.147	0.120	0.080
ENC-t, VARBS	0.097	0.100	0.107	0.100	0.107	0.099	0.104	0.100	0.106	0.107	0.103	0.095
ENC-t, FRBS	0.095	0.092	0.088	0.082	0.109	0.085	0.057	0.051	0.111	0.072	0.047	0.054
MSE-t, normal	0.001	0.003	0.010	0.038	0.000	0.005	0.014	0.034	0.001	0.004	0.011	0.027
ENC-t, normal	0.084	0.091	0.102	0.111	0.096	0.090	0.099	0.110	0.098	0.100	0.097	0.102
score-t, normal	0.001	0.004	0.013	0.032	0.002	0.008	0.013	0.022	0.002	0.003	0.010	0.021

Notes:
1. The data generating process is a set of AR(4) equations, with coefficients given in Table 1 and error variance-covariance matrix given in section 4.3.
2. For each artificial data set, point forecasts of the trivariate vector of variables are formed recursively using AR(4) and VAR(4) models estimated by OLS. Density forecasts from AR(4) and VAR(4) models are obtained (with recursive estimation) using the Bayesian MCMC estimation approach described in section 4.2, under an extremely loose (Normal-inverted Wishart) prior. At each forecast origin and horizon, we compute the log score using an empirical estimate of the forecast density obtained with a Gaussian kernel and 2000 draws of forecasts from the posterior distribution. The point forecasts and log scores are then used to form the indicated test statistics. R and P refer to the number of in-sample observations and 1-step ahead forecasts, respectively.
3. The test statistics MSE-F, MSE-t, ENC-F, ENC-t, and score-t are defined in section 4.1. All tests are conducted on a one-sided basis.
4. VARBS refers to a VAR-based bootstrap, described in section 4.1. FRBS refers to a fixed-regressor bootstrap, described in section 4.1. "Normal" refers to critical values from the standard normal distribution. The bootstraps used 499 replications. The number of Monte Carlo simulations is 2000.

Table 4. Monte Carlo Rejection Rates, 10% critical values: DGP 1, Recursive Scheme, BVAR vs. AR
(10% critical values)

	forecasts of $y_{1,t}$				forecasts of $y_{2,t}$				forecasts of $y_{3,t}$			
test, critical values	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$
R,P = 80,40												
MSE-F, VARBS	0.112	0.105	0.111	0.131	0.098	0.095	0.088	0.100	0.102	0.097	0.103	0.102
MSE-F, FRBS	0.129	0.131	0.137	0.134	0.104	0.129	0.115	0.088	0.103	0.112	0.093	0.065
MSE-t, VARBS	0.115	0.107	0.107	0.105	0.104	0.100	0.101	0.097	0.105	0.104	0.102	0.102
MSE-t, FRBS	0.123	0.117	0.099	0.120	0.104	0.104	0.069	0.073	0.104	0.085	0.055	0.071
ENC-F, VARBS	0.104	0.106	0.115	0.137	0.082	0.089	0.084	0.105	0.083	0.104	0.110	0.112
ENC-F, FRBS	0.114	0.132	0.121	0.113	0.104	0.145	0.130	0.096	0.086	0.134	0.115	0.077
ENC-t, VARBS	0.115	0.104	0.105	0.101	0.093	0.093	0.087	0.096	0.093	0.093	0.099	0.107
ENC-t, FRBS	0.108	0.095	0.088	0.094	0.096	0.083	0.058	0.073	0.088	0.068	0.051	0.070
MSE-t, normal	0.036	0.045	0.073	0.091	0.026	0.030	0.068	0.098	0.010	0.026	0.066	0.107
ENC-t, normal	0.153	0.154	0.163	0.144	0.130	0.128	0.147	0.163	0.117	0.141	0.171	0.185
score-t, normal	0.029	0.050	0.070	0.085	0.026	0.046	0.071	0.111	0.005	0.021	0.052	0.104
R,P = 80,80												
MSE-F, VARBS	0.110	0.110	0.102	0.100	0.107	0.097	0.097	0.107	0.103	0.092	0.091	0.096
MSE-F, FRBS	0.108	0.129	0.124	0.118	0.106	0.110	0.103	0.090	0.075	0.074	0.062	0.056
MSE-t, VARBS	0.113	0.113	0.100	0.091	0.105	0.096	0.102	0.106	0.099	0.093	0.092	0.096
MSE-t, FRBS	0.104	0.121	0.105	0.098	0.105	0.092	0.061	0.060	0.079	0.068	0.034	0.050
ENC-F, VARBS	0.087	0.087	0.092	0.100	0.095	0.091	0.098	0.111	0.061	0.085	0.092	0.107
ENC-F, FRBS	0.099	0.117	0.115	0.091	0.111	0.141	0.139	0.104	0.072	0.114	0.090	0.071
ENC-t, VARBS	0.097	0.101	0.091	0.096	0.100	0.095	0.101	0.110	0.075	0.085	0.096	0.109
ENC-t, FRBS	0.090	0.083	0.070	0.076	0.097	0.066	0.053	0.058	0.072	0.052	0.032	0.052
MSE-t, normal	0.011	0.018	0.040	0.050	0.005	0.011	0.036	0.068	0.002	0.009	0.029	0.065
ENC-t, normal	0.128	0.136	0.114	0.123	0.130	0.112	0.132	0.146	0.096	0.114	0.131	0.153
score-t, normal	0.013	0.027	0.038	0.059	0.009	0.018	0.042	0.058	0.002	0.009	0.026	0.066
R,P = 80,120												
MSE-F, VARBS	0.116	0.116	0.116	0.106	0.104	0.101	0.108	0.117	0.121	0.096	0.095	0.083
MSE-F, FRBS	0.102	0.127	0.134	0.134	0.091	0.097	0.089	0.075	0.080	0.067	0.050	0.046
MSE-t, VARBS	0.115	0.116	0.110	0.100	0.102	0.100	0.115	0.106	0.109	0.095	0.096	0.085
MSE-t, FRBS	0.102	0.120	0.107	0.110	0.092	0.090	0.048	0.037	0.084	0.062	0.027	0.029
ENC-F, VARBS	0.095	0.105	0.112	0.110	0.091	0.095	0.103	0.111	0.077	0.093	0.097	0.089
ENC-F, FRBS	0.105	0.135	0.134	0.102	0.107	0.137	0.128	0.091	0.093	0.121	0.090	0.062
ENC-t, VARBS	0.102	0.108	0.102	0.097	0.094	0.098	0.099	0.097	0.089	0.093	0.093	0.088
ENC-t, FRBS	0.094	0.087	0.071	0.069	0.090	0.067	0.038	0.041	0.082	0.046	0.028	0.038
MSE-t, normal	0.005	0.011	0.019	0.039	0.003	0.009	0.022	0.043	0.001	0.005	0.014	0.042
ENC-t, normal	0.131	0.133	0.119	0.115	0.127	0.111	0.112	0.120	0.113	0.113	0.120	0.134
score-t, normal	0.005	0.009	0.026	0.044	0.005	0.013	0.021	0.037	0.002	0.006	0.009	0.033

Notes:
1. The data generating process is a set of AR(4) equations, with coefficients given in Table 1 and error variance-covariance matrix given in section 4.3.
2. For each artificial data set, point forecasts of the trivariate vector of variables are formed recursively using AR(4) models estimated by OLS and a BVAR(4) estimated under a Normal-Wishart prior (using iteration and the posterior mean coefficients). Density forecasts from AR(4) and BVAR(4) models are obtained using the Bayesian MCMC estimation approach described in section 4.2, with an extremely loose prior for the AR models. At each forecast origin and horizon, we compute the log score using an empirical estimate of the forecast density obtained with a Gaussian kernel and 2000 draws of forecasts. The point forecasts and log scores are then used to form the indicated test statistics. R and P refer to the number of in-sample observations and 1-step ahead forecasts, respectively.
3. The test statistics MSE-F, MSE-t, ENC-F, ENC-t, and score-t are defined in section 4.1. All tests are conducted on a one-sided basis.
4. VARBS refers to a VAR-based bootstrap, described in section 4.1. FRBS refers to a fixed-regressor bootstrap, described in section 4.1. "Normal" refers to critical values from the standard normal distribution. The bootstraps used 499 replications. The number of Monte Carlo simulations is 2000.

Table 5. Monte Carlo Rejection Rates, 10% critical values: DGP 1, Rolling Scheme, VAR vs. AR

	forecasts of $y_{1,t}$				forecasts of $y_{2,t}$				forecasts of $y_{3,t}$			
test, critical values	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$
R,P = 80,40												
MSE-F, VARBS	0.098	0.103	0.107	0.120	0.097	0.097	0.104	0.115	0.106	0.117	0.115	0.114
MSE-F, FRBS	0.106	0.127	0.158	0.147	0.111	0.133	0.134	0.112	0.123	0.125	0.115	0.093
MSE-t, VARBS	0.098	0.102	0.105	0.103	0.097	0.095	0.108	0.100	0.106	0.111	0.109	0.096
MSE-t, FRBS	0.107	0.129	0.146	0.143	0.111	0.124	0.106	0.110	0.122	0.107	0.080	0.092
ENC-F, VARBS	0.100	0.105	0.118	0.119	0.099	0.104	0.106	0.114	0.112	0.116	0.111	0.118
ENC-F, FRBS	0.104	0.127	0.122	0.084	0.115	0.153	0.139	0.097	0.121	0.161	0.144	0.094
ENC-t, VARBS	0.092	0.103	0.107	0.103	0.096	0.098	0.096	0.105	0.113	0.114	0.110	0.108
ENC-t, FRBS	0.095	0.101	0.100	0.098	0.103	0.091	0.084	0.100	0.118	0.098	0.083	0.093
MSE-t, normal	0.009	0.015	0.040	0.069	0.004	0.015	0.036	0.072	0.009	0.024	0.046	0.072
ENC-t, normal	0.099	0.120	0.139	0.136	0.105	0.120	0.135	0.159	0.121	0.137	0.146	0.155
score-t, normal	0.008	0.015	0.036	0.045	0.009	0.017	0.044	0.080	0.005	0.020	0.043	0.064
R,P = 80,80												
MSE-F, VARBS	0.096	0.094	0.092	0.095	0.101	0.101	0.097	0.103	0.096	0.097	0.112	0.121
MSE-F, FRBS	0.095	0.118	0.143	0.164	0.110	0.125	0.113	0.115	0.100	0.088	0.092	0.102
MSE-t, VARBS	0.099	0.095	0.096	0.088	0.101	0.104	0.098	0.096	0.095	0.099	0.107	0.114
MSE-t, FRBS	0.110	0.137	0.151	0.159	0.116	0.153	0.105	0.099	0.105	0.104	0.084	0.104
ENC-F, VARBS	0.097	0.098	0.093	0.107	0.107	0.108	0.113	0.121	0.094	0.101	0.111	0.126
ENC-F, FRBS	0.108	0.116	0.102	0.069	0.128	0.161	0.137	0.101	0.110	0.149	0.127	0.093
ENC-t, VARBS	0.096	0.099	0.091	0.086	0.107	0.114	0.106	0.100	0.092	0.100	0.102	0.120
ENC-t, FRBS	0.099	0.091	0.081	0.074	0.115	0.105	0.085	0.078	0.095	0.075	0.072	0.099
MSE-t, normal	0.000	0.003	0.014	0.026	0.000	0.002	0.018	0.036	0.000	0.003	0.013	0.044
ENC-t, normal	0.096	0.096	0.097	0.104	0.106	0.119	0.122	0.132	0.093	0.105	0.122	0.149
score-t, normal	0.001	0.003	0.015	0.018	0.002	0.003	0.018	0.040	0.000	0.001	0.015	0.036
R,P = 80,120												
MSE-F, VARBS	0.092	0.085	0.091	0.106	0.097	0.091	0.097	0.100	0.104	0.097	0.097	0.104
MSE-F, FRBS	0.092	0.119	0.151	0.197	0.092	0.103	0.116	0.128	0.104	0.081	0.084	0.108
MSE-t, VARBS	0.097	0.091	0.096	0.092	0.095	0.093	0.099	0.095	0.100	0.102	0.097	0.095
MSE-t, FRBS	0.101	0.148	0.180	0.195	0.104	0.170	0.128	0.114	0.117	0.128	0.084	0.110
ENC-F, VARBS	0.093	0.092	0.106	0.109	0.092	0.097	0.113	0.121	0.102	0.109	0.110	0.121
ENC-F, FRBS	0.105	0.115	0.118	0.072	0.116	0.151	0.140	0.095	0.122	0.159	0.127	0.087
ENC-t, VARBS	0.093	0.094	0.100	0.092	0.098	0.091	0.108	0.107	0.108	0.104	0.102	0.101
ENC-t, FRBS	0.099	0.092	0.086	0.088	0.104	0.092	0.086	0.082	0.115	0.092	0.083	0.089
MSE-t, normal	0.000	0.001	0.003	0.019	0.000	0.000	0.003	0.013	0.000	0.001	0.007	0.029
ENC-t, normal	0.091	0.093	0.104	0.111	0.099	0.106	0.121	0.130	0.107	0.111	0.114	0.130
score-t, normal	0.001	0.000	0.004	0.013	0.000	0.000	0.005	0.018	0.000	0.001	0.006	0.027

Notes:
1. The data generating process is a set of AR(4) equations, with coefficients given in Table 1 and error variance-covariance matrix given in section 4.3.
2. For each artificial data set, point forecasts of the trivariate vector of variables are formed on a rolling basis (holding the estimation sample size at R observations) using AR(4) and VAR(4) models estimated by OLS. Density forecasts from AR(4) and VAR(4) models are obtained (with rolling estimation) using the Bayesian MCMC estimation approach described in section 4.2, under an extremely loose (Normal-inverted Wishart) prior. At each forecast origin and horizon, we compute the log score using an empirical estimate of the forecast density obtained with a Gaussian kernel and 2000 draws of forecasts from the posterior distribution. The point forecasts and log scores are then used to form the indicated test statistics. R and P refer to the number of in-sample observations and 1-step ahead forecasts, respectively.
3. The test statistics MSE-F, MSE-t, ENC-F, ENC-t, and score-t are defined in section 4.1. All tests are conducted on a one-sided basis.
4. VARBS refers to a VAR-based bootstrap, described in section 4.1. FRBS refers to a fixed-regressor bootstrap, described in section 4.1. "Normal" refers to critical values from the standard normal distribution. The bootstraps used 499 replications. The number of Monte Carlo simulations is 2000.

Table 6. Monte Carlo Rejection Rates, 10% critical values: DGP 1, Rolling Scheme, BVAR vs. AR

	forecasts of $y_{1,t}$				forecasts of $y_{2,t}$				forecasts of $y_{3,t}$			
test, critical values	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$
R,P = 80,40												
MSE-F, VARBS	0.097	0.101	0.107	0.120	0.102	0.101	0.105	0.117	0.098	0.103	0.103	0.111
MSE-F, FRBS	0.109	0.136	0.139	0.122	0.103	0.130	0.120	0.088	0.104	0.112	0.097	0.065
MSE-t, VARBS	0.102	0.101	0.096	0.102	0.104	0.104	0.104	0.098	0.100	0.102	0.091	0.097
MSE-t, FRBS	0.114	0.119	0.110	0.126	0.100	0.104	0.074	0.083	0.106	0.085	0.065	0.071
ENC-F, VARBS	0.102	0.105	0.117	0.128	0.093	0.096	0.106	0.119	0.103	0.107	0.116	0.121
ENC-F, FRBS	0.101	0.119	0.115	0.088	0.096	0.143	0.123	0.089	0.092	0.132	0.111	0.063
ENC-t, VARBS	0.100	0.100	0.104	0.098	0.096	0.097	0.093	0.092	0.103	0.102	0.102	0.108
ENC-t, FRBS	0.096	0.085	0.085	0.090	0.091	0.071	0.057	0.082	0.090	0.066	0.053	0.071
MSE-t, normal	0.029	0.040	0.057	0.088	0.018	0.030	0.065	0.098	0.007	0.029	0.061	0.102
ENC-t, normal	0.151	0.164	0.167	0.139	0.144	0.142	0.157	0.178	0.133	0.161	0.168	0.187
score-t, normal	0.025	0.044	0.064	0.069	0.018	0.030	0.075	0.132	0.006	0.030	0.056	0.101
R,P = 80,80												
MSE-F, VARBS	0.101	0.103	0.098	0.090	0.107	0.107	0.104	0.112	0.081	0.095	0.103	0.120
MSE-F, FRBS	0.115	0.118	0.130	0.117	0.107	0.111	0.094	0.080	0.075	0.076	0.072	0.060
MSE-t, VARBS	0.102	0.095	0.097	0.073	0.109	0.107	0.102	0.101	0.083	0.095	0.099	0.111
MSE-t, FRBS	0.115	0.118	0.114	0.118	0.103	0.102	0.062	0.065	0.082	0.075	0.046	0.067
ENC-F, VARBS	0.096	0.104	0.096	0.104	0.112	0.111	0.115	0.122	0.086	0.102	0.114	0.126
ENC-F, FRBS	0.105	0.120	0.100	0.061	0.127	0.147	0.127	0.091	0.085	0.126	0.090	0.059
ENC-t, VARBS	0.100	0.089	0.095	0.080	0.110	0.108	0.105	0.104	0.087	0.100	0.102	0.119
ENC-t, FRBS	0.094	0.080	0.071	0.065	0.099	0.075	0.054	0.066	0.081	0.052	0.046	0.063
MSE-t, normal	0.006	0.014	0.026	0.040	0.005	0.007	0.029	0.060	0.001	0.009	0.027	0.074
ENC-t, normal	0.165	0.154	0.135	0.115	0.168	0.141	0.151	0.161	0.111	0.139	0.154	0.190
score-t, normal	0.007	0.011	0.029	0.049	0.008	0.015	0.035	0.065	0.001	0.006	0.025	0.062
R,P = 80,120												
MSE-F, VARBS	0.100	0.100	0.097	0.096	0.093	0.089	0.096	0.100	0.101	0.104	0.101	0.103
MSE-F, FRBS	0.099	0.118	0.132	0.139	0.075	0.085	0.078	0.068	0.088	0.066	0.058	0.053
MSE-t, VARBS	0.101	0.100	0.088	0.088	0.092	0.087	0.093	0.096	0.114	0.103	0.095	0.098
MSE-t, FRBS	0.101	0.122	0.127	0.135	0.082	0.101	0.051	0.058	0.101	0.072	0.041	0.049
ENC-F, VARBS	0.097	0.102	0.115	0.110	0.099	0.104	0.114	0.119	0.119	0.109	0.114	0.118
ENC-F, FRBS	0.101	0.119	0.111	0.060	0.115	0.142	0.116	0.074	0.113	0.115	0.089	0.046
ENC-t, VARBS	0.099	0.092	0.099	0.090	0.107	0.094	0.099	0.103	0.107	0.107	0.111	0.111
ENC-t, FRBS	0.095	0.074	0.068	0.070	0.097	0.070	0.046	0.051	0.099	0.052	0.043	0.051
MSE-t, normal	0.002	0.005	0.010	0.025	0.001	0.002	0.009	0.028	0.000	0.004	0.013	0.049
ENC-t, normal	0.177	0.167	0.145	0.116	0.173	0.138	0.146	0.153	0.141	0.167	0.158	0.185
score-t, normal	0.004	0.006	0.014	0.030	0.002	0.004	0.018	0.027	0.000	0.004	0.009	0.052

Notes:
1. The data generating process is a VAR(4), with coefficients given in Table 2 and error variance-covariance matrix given in section 4.3.
2. For each artificial data set, point forecasts of the trivariate vector of variables are formed on a rolling basis (holding the estimation sample size at R observations) using AR(4) models estimated by OLS and a BVAR(4) estimated with a Normal-inverted Wishart prior and posterior (using iteration and the posterior mean coefficients). Density forecasts from AR(4) and BVAR(4) models are obtained (with rolling estimation) using the Bayesian MCMC estimation approach described in section 4.2, using an extremely loose (Normal-inverted Wishart) prior for the AR models. At each forecast origin and horizon, we compute the log score using an empirical estimate of the forecast density obtained with a Gaussian kernel and 2000 draws of forecasts from the posterior distribution. The point forecasts and log scores are then used to form the indicated test statistics. R and P refer to the number of in-sample observations and 1-step ahead forecasts, respectively.
3. The test statistics MSE-F, MSE-t, ENC-F, ENC-t, and score-t are defined in section 4.1. All tests are conducted on a one-sided basis.
4. VARBS refers to a VAR-based bootstrap, described in section 4.1. FRBS refers to a fixed-regressor bootstrap, described in section 4.1. "Normal" refers to critical values from the standard normal distribution. The bootstraps used 499 replications. The number of Monte Carlo simulations is 2000.

Table 7. Monte Carlo Rejection Rates, 10% critical values: DGP 2, Recursive Scheme, VAR vs. AR
(10% critical values)

	forecasts of $y_{1,t}$				forecasts of $y_{2,t}$				forecasts of $y_{3,t}$			
test, critical values	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$
R,P = 80,40												
MSE-F, VARBS	0.708	0.707	0.489	0.432	0.636	0.488	0.414	0.220	0.646	0.681	0.519	0.415
MSE-F, FRBS	0.710	0.736	0.515	0.462	0.634	0.537	0.456	0.249	0.660	0.680	0.490	0.380
MSE-t, VARBS	0.690	0.664	0.392	0.263	0.606	0.454	0.328	0.162	0.623	0.625	0.420	0.272
MSE-t, FRBS	0.676	0.676	0.373	0.292	0.601	0.475	0.284	0.159	0.631	0.476	0.193	0.180
ENC-F, VARBS	0.889	0.875	0.598	0.517	0.786	0.610	0.472	0.235	0.834	0.862	0.645	0.509
ENC-F, FRBS	0.913	0.910	0.645	0.502	0.828	0.737	0.599	0.280	0.870	0.913	0.726	0.532
ENC-t, VARBS	0.835	0.796	0.425	0.282	0.720	0.566	0.381	0.172	0.778	0.780	0.528	0.328
ENC-t, FRBS	0.839	0.784	0.349	0.256	0.735	0.554	0.355	0.173	0.789	0.683	0.372	0.270
MSE-t, normal	0.263	0.322	0.207	0.202	0.217	0.188	0.182	0.130	0.234	0.308	0.255	0.228
ENC-t, normal	0.841	0.823	0.507	0.364	0.740	0.613	0.462	0.257	0.798	0.814	0.633	0.469
score-t, normal	0.285	0.340	0.250	0.249	0.213	0.202	0.214	0.146	0.210	0.293	0.261	0.290
R,P = 80,80												
MSE-F, VARBS	0.901	0.903	0.680	0.583	0.826	0.689	0.573	0.301	0.846	0.868	0.705	0.578
MSE-F, FRBS	0.893	0.904	0.692	0.612	0.814	0.699	0.589	0.327	0.843	0.822	0.625	0.489
MSE-t, VARBS	0.918	0.909	0.632	0.445	0.841	0.683	0.531	0.267	0.862	0.866	0.649	0.470
MSE-t, FRBS	0.905	0.903	0.589	0.441	0.826	0.671	0.433	0.221	0.863	0.685	0.236	0.210
ENC-F, VARBS	0.983	0.983	0.796	0.700	0.944	0.830	0.688	0.335	0.969	0.973	0.833	0.692
ENC-F, FRBS	0.988	0.990	0.840	0.693	0.957	0.890	0.765	0.389	0.975	0.987	0.871	0.684
ENC-t, VARBS	0.976	0.970	0.708	0.486	0.928	0.812	0.642	0.288	0.955	0.951	0.779	0.577
ENC-t, FRBS	0.974	0.967	0.606	0.405	0.931	0.790	0.581	0.273	0.956	0.899	0.547	0.400
MSE-t, normal	0.393	0.447	0.252	0.255	0.294	0.230	0.202	0.118	0.298	0.410	0.302	0.270
ENC-t, normal	0.975	0.972	0.725	0.556	0.928	0.816	0.666	0.349	0.954	0.957	0.803	0.647
score-t, normal	0.376	0.448	0.339	0.333	0.285	0.229	0.227	0.118	0.283	0.373	0.288	0.296
R,P = 80,120												
MSE-F, VARBS	0.971	0.968	0.793	0.706	0.928	0.824	0.689	0.372	0.938	0.957	0.821	0.691
MSE-F, FRBS	0.965	0.966	0.788	0.724	0.915	0.801	0.673	0.365	0.937	0.926	0.722	0.561
MSE-t, VARBS	0.979	0.978	0.780	0.612	0.944	0.834	0.675	0.342	0.955	0.963	0.804	0.622
MSE-t, FRBS	0.976	0.979	0.727	0.573	0.938	0.819	0.535	0.263	0.957	0.840	0.267	0.242
ENC-F, VARBS	0.997	0.997	0.911	0.828	0.986	0.923	0.800	0.409	0.994	0.995	0.939	0.819
ENC-F, FRBS	0.999	0.997	0.932	0.818	0.988	0.957	0.863	0.449	0.995	0.996	0.939	0.780
ENC-t, VARBS	0.994	0.996	0.868	0.679	0.981	0.920	0.780	0.390	0.992	0.992	0.915	0.772
ENC-t, FRBS	0.995	0.995	0.796	0.543	0.984	0.906	0.722	0.332	0.991	0.979	0.711	0.530
MSE-t, normal	0.532	0.597	0.344	0.327	0.409	0.290	0.245	0.139	0.429	0.554	0.383	0.326
ENC-t, normal	0.995	0.995	0.865	0.715	0.980	0.919	0.785	0.416	0.991	0.992	0.918	0.796
score-t, normal	0.506	0.589	0.419	0.417	0.386	0.297	0.277	0.137	0.393	0.515	0.400	0.370

Notes:
1. The data generating process is VAR(4), with coefficients given in Table 2 and error variance-covariance matrix given in section 4.3.
2. For each artificial data set, point forecasts of the trivariate vector of variables are formed recursively using AR(4) and VAR(4) models estimated by OLS. Density forecasts from AR(4) and VAR(4) models are obtained (with recursive estimation) using the Bayesian MCMC estimation approach described in section 4.2, under an extremely loose (Normal-inverted Wishart) prior. At each forecast origin and horizon, we compute the log score using an empirical estimate of the forecast density obtained with a Gaussian kernel and 2000 draws of forecasts from the posterior distribution. The point forecasts and log scores are then used to form the indicated test statistics. R and P refer to the number of in-sample observations and 1-step ahead forecasts, respectively.
3. The test statistics MSE-F, MSE-t, ENC-F, ENC-t, and score-t are defined in section 4.1. All tests are conducted on a one-sided basis.
4. VARBS refers to a VAR-based bootstrap, described in section 4.1. FRBS refers to a fixed-regressor bootstrap, described in section 4.1. "Normal" refers to critical values from the standard normal distribution. The bootstraps used 499 replications. The number of Monte Carlo simulations is 2000.

Table 8. Monte Carlo Rejection Rates, 10% critical values: DGP 2, Recursive Scheme, BVAR vs. AR
(10% critical values)

	forecasts of $y_{1,t}$				forecasts of $y_{2,t}$				forecasts of $y_{3,t}$			
test, critical values	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$
R,P = 80,40												
MSE-F, VARBS	0.740	0.742	0.512	0.457	0.627	0.457	0.371	0.184	0.610	0.684	0.507	0.401
MSE-F, FRBS	0.749	0.774	0.543	0.464	0.650	0.535	0.453	0.220	0.598	0.678	0.466	0.338
MSE-t, VARBS	0.664	0.625	0.369	0.237	0.579	0.419	0.307	0.151	0.596	0.599	0.392	0.251
MSE-t, FRBS	0.681	0.640	0.303	0.227	0.591	0.435	0.272	0.145	0.570	0.410	0.124	0.148
ENC-F, VARBS	0.889	0.889	0.619	0.547	0.692	0.460	0.355	0.165	0.762	0.837	0.645	0.485
ENC-F, FRBS	0.914	0.917	0.667	0.542	0.743	0.619	0.512	0.225	0.805	0.894	0.684	0.451
ENC-t, VARBS	0.808	0.769	0.398	0.275	0.635	0.456	0.317	0.149	0.713	0.720	0.468	0.290
ENC-t, FRBS	0.816	0.745	0.279	0.219	0.649	0.430	0.269	0.144	0.725	0.527	0.192	0.181
MSE-t, normal	0.387	0.416	0.258	0.206	0.289	0.235	0.228	0.157	0.188	0.340	0.289	0.266
ENC-t, normal	0.864	0.851	0.541	0.366	0.712	0.538	0.439	0.245	0.767	0.807	0.653	0.502
score-t, normal	0.415	0.458	0.313	0.281	0.317	0.262	0.270	0.189	0.163	0.333	0.285	0.341
R,P = 80,80												
MSE-F, VARBS	0.913	0.917	0.719	0.596	0.842	0.695	0.573	0.273	0.800	0.867	0.707	0.569
MSE-F, FRBS	0.919	0.924	0.721	0.606	0.840	0.719	0.602	0.299	0.767	0.832	0.579	0.439
MSE-t, VARBS	0.897	0.887	0.618	0.434	0.831	0.682	0.526	0.249	0.814	0.843	0.624	0.437
MSE-t, FRBS	0.899	0.884	0.494	0.342	0.829	0.666	0.416	0.177	0.778	0.592	0.127	0.156
ENC-F, VARBS	0.982	0.984	0.817	0.715	0.905	0.706	0.567	0.237	0.926	0.962	0.830	0.659
ENC-F, FRBS	0.988	0.990	0.861	0.709	0.923	0.813	0.709	0.328	0.942	0.981	0.842	0.601
ENC-t, VARBS	0.975	0.963	0.693	0.492	0.882	0.730	0.562	0.252	0.909	0.929	0.743	0.527
ENC-t, FRBS	0.974	0.960	0.516	0.340	0.888	0.694	0.471	0.216	0.916	0.791	0.271	0.249
MSE-t, normal	0.547	0.589	0.326	0.272	0.447	0.308	0.271	0.161	0.240	0.445	0.345	0.329
ENC-t, normal	0.981	0.980	0.763	0.572	0.909	0.766	0.635	0.328	0.928	0.955	0.832	0.669
score-t, normal	0.538	0.595	0.414	0.375	0.452	0.327	0.302	0.138	0.239	0.443	0.330	0.354
R,P = 80,120												
MSE-F, VARBS	0.977	0.976	0.818	0.727	0.942	0.829	0.705	0.351	0.914	0.960	0.826	0.683
MSE-F, FRBS	0.976	0.976	0.816	0.726	0.937	0.824	0.699	0.341	0.894	0.925	0.668	0.516
MSE-t, VARBS	0.976	0.970	0.770	0.605	0.943	0.822	0.680	0.326	0.939	0.960	0.788	0.586
MSE-t, FRBS	0.977	0.968	0.653	0.475	0.933	0.802	0.524	0.220	0.919	0.733	0.125	0.163
ENC-F, VARBS	0.998	0.998	0.920	0.843	0.967	0.844	0.699	0.315	0.981	0.994	0.934	0.791
ENC-F, FRBS	0.999	0.998	0.939	0.839	0.977	0.910	0.808	0.376	0.985	0.996	0.923	0.720
ENC-t, VARBS	0.994	0.996	0.866	0.670	0.961	0.857	0.711	0.332	0.976	0.984	0.901	0.713
ENC-t, FRBS	0.995	0.994	0.729	0.466	0.967	0.824	0.612	0.260	0.979	0.923	0.372	0.341
MSE-t, normal	0.685	0.709	0.433	0.356	0.572	0.402	0.337	0.161	0.331	0.602	0.416	0.395
ENC-t, normal	0.997	0.997	0.885	0.724	0.972	0.873	0.745	0.381	0.982	0.990	0.930	0.809
score-t, normal	0.673	0.722	0.521	0.447	0.572	0.426	0.370	0.169	0.342	0.584	0.432	0.423

Notes:
1. The data generating process is a VAR(4), with coefficients given in Table 2 and error variance-covariance matrix given in section 4.3.
2. For each artificial data set, point forecasts of the trivariate vector of variables are formed recursively using AR(4) models estimated by OLS and a BVAR(4) estimated under a Normal-Wishart prior (using iteration and the posterior mean coefficients). Density forecasts from AR(4) and BVAR(4) models are obtained using the Bayesian MCMC estimation approach described in section 4.2, with an extremely loose prior for the AR models. At each forecast origin and horizon, we compute the log score using an empirical estimate of the forecast density obtained with a Gaussian kernel and 2000 draws of forecasts. The point forecasts and log scores are then used to form the indicated test statistics. R and P refer to the number of in-sample observations and 1-step ahead forecasts, respectively.
3. The test statistics MSE-F, MSE-t, ENC-F, ENC-t, and score-t are defined in section 4.1. All tests are conducted on a one-sided basis.
4. VARBS refers to a VAR-based bootstrap, described in section 4.1. FRBS refers to a fixed-regressor bootstrap, described in section 4.1. "Normal" refers to critical values from the standard normal distribution. The bootstraps used 499 replications. The number of Monte Carlo simulations is 2000.