



Research Division
Federal Reserve Bank of St. Louis
Working Paper Series



Averaging Forecasts from VARs with Uncertain Instabilities

Todd E. Clark
and
Michael W. McCracken

Working Paper 2008-030B
<http://research.stlouisfed.org/wp/2008/2008-030.pdf>

August 2008
Revised October 2009

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
P.O. Box 442
St. Louis, MO 63166

The views expressed are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

Federal Reserve Bank of St. Louis Working Papers are preliminary materials circulated to stimulate discussion and critical comment. References in publications to Federal Reserve Bank of St. Louis Working Papers (other than an acknowledgment that the writer has had access to unpublished material) should be cleared with the author or authors.

Averaging Forecasts from VARs with Uncertain Instabilities *

Todd E. Clark
Federal Reserve Bank of Kansas City

Michael W. McCracken
Board of Governors of the Federal Reserve System

May 2008

Abstract

Recent work suggests VAR models of output, inflation, and interest rates may be prone to instabilities. In the face of such instabilities, a variety of estimation or forecasting methods might be used to improve the accuracy of forecasts from a VAR. The uncertainty inherent in any single representation of instability could mean that combining forecasts from a range of approaches will improve forecast accuracy. Focusing on models of U.S. output, prices, and interest rates, this paper examines the effectiveness of combining various models of instability in improving VAR forecasts made with real-time data.

JEL Nos.: C53, E37, C32

Keywords: Forecast combination, real-time data, structural change

* *Clark (corresponding author)*: Economic Research Dept., Federal Reserve Bank of Kansas City; 925 Grand; Kansas City, MO 64198; todd.e.clark@kc.frb.org; phone: (816)881-2575; fax: (816)881-2199. *McCracken*: Board of Governors of the Federal Reserve System; 20th and Constitution N.W.; Mail Stop #61; Washington, D.C. 20551; michael.w.mccracken@frb.gov. This paper was written for a Reserve Bank of New Zealand conference Macroeconometrics and Model Uncertainty held in June 2006. We gratefully acknowledge helpful conversations with Simon Potter and Shaun Vahey and helpful comments from Christie Smith, other conference participants, and two referees. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City, Board of Governors, Federal Reserve System, or any of its staff.

1 Introduction

Small-scale VARs are now widely used in forecasting (see, e.g., Jacobson et al. (2001), Robertson and Tallman (2001), Del Negro and Schorfheide (2004), and Favero and Marcellino (2005)). However, there is an increasing body of evidence suggesting that these VARs may be prone to instabilities (see, e.g., Kozicki and Tinsley (2001, 2002), Cogley and Sargent (2005), and Boivin (2006)). Although many different structural forces could lead to instabilities in macroeconomic VARs, much of the aforementioned literature has focused on shifts potentially attributable to changes in the behavior of monetary policy.

Accordingly, Clark and McCracken (2006) consider various methods for improving the forecast accuracy of VARs in the presence of structural change, including: sequentially updating lag orders, using various observation windows for estimation, working in differences rather than levels, making intercept corrections, allowing stochastic time variation in model parameters, allowing discrete breaks in parameters, discounted least squares estimation, Bayesian shrinkage, and detrending of inflation and interest rates. Simple averages (across the various methods just described) were consistently among the best performers.

Our preferred interpretation of this result is that in practice it is very difficult to know the form of structural instability, and model averaging provides an effective method for forecasting in the face of such uncertainty. As summarized by Timmermann (2006), competing models will differ in their sensitivity to structural breaks. Depending on the size and nature of structural breaks, models that quickly pick up changes in parameters may or may not be more accurate than models that do not. For instance, in the case of a small, recent break, a model with constant parameters may forecast more accurately than a model that allows a break in coefficients, due to the additional noise introduced by the estimation of post-break coefficients (see, for example, Clark and McCracken (2008) and Pesaran and Timmermann (2007)). However, in the case of a large break well in the past, a model that correctly picks up the associated change in coefficients will likely forecast more accurately than models with constant or slowly changing parameters. Accordingly, Timmermann (2006) and Pesaran and Timmermann (2007) suggest that combinations of forecasts from models with varying

degrees of adaptability to uncertain structural breaks will be more accurate than forecasts from individual models.

This paper provides evidence on the ability of various forms of forecast averaging to improve the real-time forecast accuracy of small-scale macroeconomic VARs in the presence of uncertain forms of model instabilities. We consider a wide range of approaches to averaging forecasts obtained with a variety of the aforementioned primitive methods for managing model instability. The average forecasts include: equally weighted averages with and without trimming, medians, common factor-based forecasts, Bates–Granger combinations estimated with ridge regression, MSE-weighted averages, lowest MSE (predictive least squares) forecasts, Bayesian model averages, and combinations based on quartile average forecasts.¹ For each of these averaging approaches, we construct real time forecasts of each variable using real-time data. We compare our results to those from simple baseline univariate models and selected baseline VAR models.

Our results indicate that while some of the primitive forms of managing structural instability sometimes provide the largest gains in terms of forecast accuracy — notably those models with some form of Bayesian shrinkage — model averaging is a more consistent method for improving forecast accuracy. Not surprisingly, the best type of averaging often varies with the variable being forecast. However, after aggregating across all models, horizons and variables, it is clear that the simplest forms of model averaging — such as those that use equal weights across all models — consistently perform among the best methods. The best forecast is a simple average of projections from a univariate model and a VAR using using detrended inflation and interest rates. At the other extreme, forecasts based on OLS-type combination and factor model-based combination rank among the worst.

The remainder of the paper proceeds as follows. Section 2 describes the real-time data and samples. Section 3 provides a synopsis of the forms of model averaging used to forecast.

¹In the interest of tractability, we abstract from the predictive likelihood approach to model and forecast averaging examined in Eklund and Karlsson (2007). Our forecasts based on out-of-sample MSE weighting of course incorporate some of the flavor of weighting on the basis of predictive likelihood. However, the computational intensity of our analysis (large numbers of variables and models, long samples, etc.) and the variety of approaches used to model estimation (OLS for full and rolling samples, Minneapolis BVAR estimation for full and rolling samples, etc.) would pose challenges to the inclusion of a predictive likelihood approach. Accordingly, we leave further analysis of predictive likelihood approaches to future research.

Section 4 presents our results on forecast accuracy. Section 5 concludes.

2 Data

We consider the real-time forecast performance of models with three different measures of output (y), two measures of inflation (π), and a short-term interest rate (i). The output measures are GDP or GNP (depending on data vintage) growth, an output gap (hereafter, the HPS gap) computed in real time with the method described in Hallman, et al. (1991), and an output gap estimated in real time with the HP filter. The inflation measures include the GDP or GNP deflator or price index (depending on data vintage) and CPI. The interest rate is the 3-month Treasury bill rate; using the federal funds rate yields qualitatively similar results. Growth and inflation rates are measured as annualized log changes (from $t - 1$ to t). Output gaps are measured in percentages (100 times the log of output relative to trend). Interest rates are expressed in annualized percentage points.

The raw data are from the Federal Reserve Bank of Philadelphia’s Real-Time Data Set for Macroeconomists (RTDSM), the Board of Governor’s FAME database, and the website of the Bureau of Labor Statistics (BLS). Real-time data on GDP or GNP and the GDP or GNP price series are from the RTDSM. Hereafter we simply use the notation “GDP” and “GDP price index” to refer to the output and price series, even though the measures are based on GNP and a fixed weight deflator for much of the sample. In the case of the CPI and the interest rates, for which real time revisions are small to essentially non-existent, we simply abstract from real time aspects of the data. For the CPI, we follow the advice of Kozicki and Hoffman (2004) for avoiding choppiness in inflation rates for the 1960s and 1970s due to changes in index bases, and use a 1967 base year series taken from the BLS website. For the T-bill rate, we use a series obtained from FAME.

The full forecast evaluation period runs from 1970:Q1 through 2005; as detailed in section 3, forecasts from 1965:Q4 through 1969:Q4 are used as initial values in the combination forecasts that require historical forecasts. Accordingly, we use real time data vintages from 1965:Q4 through 2005:Q4. The vintages of the RTDSM are dated to reflect the information available around the middle of each quarter. Normally, in vintage t , the available NIPA data

run through period $t - 1$. The start dates of the raw data available in each vintage vary over time, ranging from 1947:Q1 to 1959:Q3, reflecting changes in the published samples of the historical data. At each forecast origin t , we use vintage t data to estimate output gaps and the forecast models and then construct forecasts for periods t and beyond. The starting point of the model estimation sample is the maximum of (i) 1955:Q1 and (ii) the earliest quarter in which all of the data in a given model are available, plus five quarters to allow for four lags and differencing or detrending.

We examine accuracy results for forecast horizons of the current quarter ($h = 0Q$), the next quarter ($h = 1Q$), four quarters ahead ($h = 1Y$), and eight quarters ahead ($h = 2Y$). In keeping with common central bank practice, the 1- and 2-year ahead forecasts for GDP/GNP growth and inflation are four-quarter rates of change. The 1- and 2-year ahead forecasts for output gaps and interest rates are quarterly levels in periods $t + 4$ and $t + 8$, respectively. All of the multi-step forecasts are obtained by iterating the 1-step ahead models.

We follow Romer and Romer (2000) and use the second available estimates of GDP/GNP and the GDP/GNP deflator as actuals in evaluating forecast accuracy.² In the case of h -step ahead forecasts made for period $t + h$ with vintage t data ending in period $t - 1$, the second available estimate is normally taken from the vintage $t + h + 2$ data set. In light of our abstraction from real time revisions in CPI inflation and interest rates, for these series the real time data correspond to the final vintage data.

3 Forecast methods

The forecasts of interest in this paper are combinations of forecasts from a wide range of approaches to allowing for structural change in trivariate VARs. Table 1 lists the set of individual VAR forecast methods considered in this paper, along with some detail on forecast construction. To be precise, for each model — defined as being a baseline VAR in one measure of output (y), one measure of inflation (π), and one short-term interest rate (i) — we apply each of the estimation and forecasting methods listed in Table 1.

Note that, although we simply refer to all the underlying forecasts as VAR forecasts,

²Our broad findings are highly robust to alternative definitions of actuals: 1st available, 5th available, and final vintage.

in fact the list of individual models includes a univariate specification for each measure of output, inflation, and the interest rate. For output the univariate model is an AR(2). In the case of inflation, we follow Stock and Watson (2007) and use an MA(1) process for the change in inflation ($\Delta\pi$), estimated with a rolling window of 40 observations. The univariate model for the short-term interest rate is also specified as a rolling MA(1) in the first difference of the series (Δi).

Table 2 details all of the approaches we use to combining forecasts from these underlying models. The remainder of this section explains the averaging methods.

3.1 Equally weighted averages

We begin with seven simple forms of model averaging, each using what could loosely be described as equal weights. The first is an equally weighted average of all the VAR forecasts in Table 1. Specifically, for a given combination of measures of output, inflation, and the interest rate (for example, for the combination GDP growth, GDP inflation, and the T-bill rate), we average forecasts from the 50 VARs listed in Table 1. We also consider the median forecast and 10 and 20 percent trimmed means.

We include a fifth average forecast approach motivated by Clark and McCracken (2008), who show that forecast accuracy can be improved by combining forecasts from models estimated with recursive (all available data) and rolling samples. For a given VAR(4), we form an equally weighted average of the model forecasts constructed using parameters estimated (i) recursively and (ii) with a rolling window of the past 60 observations. Three other averages are motivated by the Clark and McCracken (2005a) finding that combining forecasts from nested models can improve forecast accuracy. We consider an average of the univariate forecast with the VAR(4) forecast, an average of the univariate forecast with the DVAR(4) forecast, and an average of the univariate forecast with a forecast from a VAR(4) in output, detrended inflation, and the detrended interest rate (see Table 1 and section 3.7 for more information on the detrended VAR).

3.2 Combinations based on Bates–Granger/ridge regression

We also consider a large number of average forecasts based on historical forecast performance — one such approach being forecast combination based on Bates and Granger (1969) regression. For these methods, we need an initial sample of forecasts preceding the sample to be used in our formal forecast evaluation. We use an initial sample of forecasts from 1965:Q4 (the starting point of the RTDSM) through 1969:Q4. Therefore, in the case of current quarter forecasts constructed in 1970:Q1, we have an initial sample of 17 forecasts to use in estimating combination regressions, forming MSE weights, etc.

To obtain Bates–Granger combinations, for each of output, inflation, and the interest rate we use the data that would have been available to a forecaster in real time to estimate a generalized ridge regression of the actual data on the 50 VAR forecasts, shrinking the coefficients toward equal weights. Our implementation follows Stock and Watson (1999): letting $Z_{t+h|t}$ denote the vector of 50 forecasts of variable z_{t+h} made in period t and β^{equal} denote a 50×1 vector filled with $1/50$, the combination coefficient vector estimate is

$$\hat{\beta} = (cI_{50} + \sum_t Z_{t+h|t}Z'_{t+h|t})^{-1}(c\beta^{equal} + \sum_t Z_{t+h|t}z_{t+h}), \quad (1)$$

where $c = k \times \text{trace}(50^{-1} \sum_t Z_{t+h|t}Z'_{t+h|t})$. We consider three different forecasts, based on different values of the shrinkage coefficient k : .001, .25, and 1. A smaller (larger) value of k implies less (more) shrinkage. We use $k = .001$ to approximate OLS combination. For each k , we consider forecasts based on both a recursive estimate of the combination regression and a 10–year rolling sample estimate (using all available if the sample is shorter than 10 years).

3.3 Common factor combinations

Stock and Watson (1999, 2004) develop another approach to combining information from individual model forecasts: estimating a common factor from the forecasts, regressing actual data on the common factor, and then using the fitted regression to forecast into the future. Therefore, using the real time forecasts available through the forecast origin t , we estimate (by principal components) one common factor from the set of 50 VAR forecasts for each of output, inflation, and the interest rate (estimating one factor for output, another for inflation,

etc.). We then regress the actual data available in real time as of t on a constant and the factor. The factor-based forecast is then obtained from the estimated regression, using the factor observation for period t .

3.4 MSE-weighted and PLS forecasts

We also consider several average forecasts based on inverse MSE weights. At each forecast origin t , historical MSEs of the 50 VAR forecasts of each of output, inflation, and the interest rate are calculated with the available forecasts and actual data, and each forecast i of the given variable is given a weight of $MSE_i^{-1} / \sum_i MSE_i^{-1}$. In addition, following Stock and Watson (2004), we consider a forecast based on a discounted mean square forecast error (in which, from a forecast origin of t , the squared error in the earlier period s is discounted by a factor δ^{t-s}). We use a discount rate of $\delta = .95$.

In addition, we consider a predictive least squares (PLS) forecast. At each forecast origin t , we identify the model forecast with the lowest historical MSE, and then use that single model to forecast. We compute alternative MSE-weighted and PLS forecasts with not only recursive and 10-year rolling samples but also a 5-year rolling sample of forecasts.

3.5 Quartile forecasts

Aiolfi and Timmermann (2006) develop alternative approaches to forecast combination that take into account persistence in forecast performance — the possibility that some models may be consistently good while others may be consistently bad. Their simplest forecast is an equally weighted average of the forecasts in the top quartile of forecast accuracy (that is, the forecasts with historical MSEs in the lowest quartile of MSEs). More sophisticated forecasts involve measuring performance persistence as forecasting moves forward in time, sorting the forecasts into clusters based on past performance, and estimating combination regressions with a number of clusters determined by the degree of persistence. For tractability, we consider simple versions of the Aiolfi–Timmermann methods, based on just the first and second quartiles. Specifically, we consider a simple average of the forecasts in the top quartile of historical forecast accuracy. We also consider a forecast based on an OLS-estimated combination regression including a constant, the average of the first quartile forecasts, and

the average of the second quartile forecasts.

3.6 Bayesian model averages

We also consider forecasts obtained by Bayesian model averaging (BMA). At each forecast origin t , for each equation of the 50 models listed in Table 1, we calculate a posterior probability from prior probabilities and marginal likelihoods for each model, with each model assigned the same prior probability.

We consider three different measures of the marginal likelihood, each of which yields a different BMA forecast: AIC, BIC, and Phillips' (1996) PIC.³ The BIC is well known to be proportional to the marginal likelihood of models estimated by OLS or, equivalently, diffuse priors.⁴ The AIC can be viewed as another measure of the marginal likelihood for models estimated by OLS.⁵ Phillips (1996) develops another criterion, PIC, as a measure of marginal likelihood appropriate for comparing VARs in levels, differences, and with informative priors.

Specifically, at each forecast origin t , for each of the model estimates listed in Table 1, we use in-sample model residuals to compute the AIC, BIC, and PIC for each equation of the model.⁶ For each criterion, we then form a BMA forecast using $-5T$ times the information criterion value as the marginal likelihood of each equation.

In our application, calculating the information criteria requires some decisions on how to deal with some of the important differences in estimation approaches (e.g., rolling versus recursive estimation) for the 50 underlying model forecasts. In the case of models estimated with a rolling sample of data, we calculate the AIC, BIC, and PIC based on a model that allows a discrete break in all the model coefficients at the point of the beginning of the rolling sample. For models estimated by discounted least squares (DLS), we calculate the information criteria using residuals defined as actual data less fitted values based on the DLS

³Our BMA forecasts are numerically equivalent to those that would be obtained under the information criteria-weighting approach, based on in-sample sums of square errors, developed in Kapetanios, et al. (2007).

⁴BMA applications such as Garratt, Koop, and Vahey (2006) have also used BIC to estimate the marginal likelihood and in turn average models.

⁵We compute the AIC without any small-sample corrections, as $T \cdot \log \hat{\sigma}^2 + 2k$, where T is the total sample size at the forecast origin, $\hat{\sigma}^2$ is the residual variance (normalized by T), and k denotes the number of regressors in the equation. We also compute the BIC and PIC without small-sample corrections.

⁶In calculating PIC for the univariate IMA models for inflation and interest rates, we simply approximate the MA fits with AR(1) models estimated for $\Delta\pi$ and Δi (estimating separate models for the rolling sample and the earlier sample), and calculate PIC values using these AR(1) approximations.

coefficient estimates.

In the case of the AIC and BIC applied to BVAR models, for simplicity we abstract from the prior and calculate the criteria based on the residual sums of squares and simple parameter count (PIC is calculated for VARs and BVARs, to take account of priors, as described in Phillips (1996)).⁷ As Phillips (1996) notes, the prior is asymptotically irrelevant in the sense that, as the sample grows, sample information dominates the prior. For marginal likelihood measures other than PIC, taking (proper Bayesian) account of the finite-sample role of the Bayesian prior in combining forecasts from models estimated with different priors would require Monte Carlo integration, which is intractable in our large-scale, real-time forecast evaluation.

3.7 Benchmark forecasts

To evaluate the practical merit of the averaging methods described above, we compare the accuracy of the above combination or average forecasts against various benchmarks. In light of common practice in forecasting research, we use forecasts from the univariate time series models as one set of benchmarks.⁸

We also include for comparison forecasts from selected VAR methods that are either of general interest in light of common usage or performed relatively well in our prior work: a VAR(4); DVAR(4) (a VAR with inflation and the interest rate differenced); BVAR(4) with conventional Minnesota priors; BVAR(4) with stochastically time-varying (random walk) parameters; and a BVAR(4) in output, detrended inflation, and the interest rate less the inflation trend. The BVAR(4) with inflation detrending draws on the work of Kozicki and Tinsley (2001) on models with learning about an unobserved time-varying inflation target of the central bank. For tractability in real time forecasting, we follow Cogley (2002) in estimating the inflation target or trend with exponential smoothing, with a smoothing parameter of .05.⁹ Table 1 provides additional detail on all of these model specifications.

⁷For BVARs with TVP, at each point in time t we calculate the model residuals as a function of the period t coefficients and use these residuals to compute the residual sums of squares.

⁸On average across horizons and samples, the univariate benchmarks we use are at least as accurate as others that might be considered reasonable, such as random walks, recursive and rolling AR models with BIC-determined lag orders, and recursive and rolling VAR models with BIC-determined lag orders.

⁹We set the smoothing parameter at .05, to resemble survey measures of long-run inflation expectations.

4 Results

In evaluating the performance of the forecasting methods described above, we use root mean square error (RMSE) to evaluate accuracy. In light of the potential for instabilities in forecast performance, we examine accuracy over forecast samples of 1970-84 and 1985-2005.¹⁰

To be able to provide broad, robust results, in total we consider a large number of models and methods — too many to be able to present all details of the results. We present more detailed results on forecasts of GDP growth and inflation than forecasts of the output gap measures or interest rates. We also focus on a few forecast horizons — those for $h = 0Q$, $h = 1Q$, and $h = 1Y$ — and present just summary results for the $h = 2Y$ horizon

Tables 3 and 4 report forecast accuracy (RMSE) results for GDP growth and either GDP price index-based or CPI-based inflation using 38 forecast methods. In each case we use the 3-month T-bill as the interest rate, and present results for horizons $h = 0Q$, $h = 1Q$, and $h = 1Y$ (results for the $h = 2Y$ horizon are available in the working paper version of this paper). In Table 5 we report forecast accuracy results for the T-bill rate, from models using GDP growth and GDP inflation. In every case, the first row of the table provides the RMSE associated with the baseline univariate model, while the others report ratios of the corresponding RMSE to that for the benchmark univariate model.

In Table 6 we take another approach to broadly determining which methods tend to perform better than the benchmark. Across each variable, model and forecast horizon, we compute the average rank of the methods included in Tables 3-5. We present average rankings for every method we consider across each variable, forecast horizon, and the 1970-84 and 1985-05 samples (spanning all columns of Tables 3-5 plus unreported results for forecasts from models using an output gap, forecasts of the T-bill rate from models using our various measures of output and inflation, and forecasts for the $h = 2Y$ horizon).

To determine the statistical significance of differences in forecast accuracy, as a rough guide we use a non-parametric bootstrap patterned after White's (2000).¹¹ The individual

¹⁰With forecasts dated by the end period of the forecast horizon $h = 0, 1, 4, 8$, the VAR forecast samples are, respectively, 1970:Q1+ h to 1984:Q4 and 1985:Q1 to 2005:Q3- h .

¹¹We say "rough guide" for two reasons. First, the models under consideration include both nested and non-nested models. The inclusion of nested models violates the technical assumptions of White (2000) and Hansen

p -values represent a pairwise comparison of each VAR or average forecast to the univariate forecast. RMSE ratios that are significantly less than 1 at a 10 percent confidence interval are indicated with a *slanted* font. To determine whether a best forecast in each column of Tables 3-5 is significantly better than the benchmark once the data snooping or search involved in selecting a best forecast is taken into account, we apply Hansen’s (2005) SPA variant of White’s (2000) reality check test to differences in MSEs (for each model relative to the benchmark). For each column, if the SPA test yields a p -value of 10 percent or less, we report the associated RMSE ratio in bold font.¹² We implement the bootstrap by sampling (with moving block methods) from the time series of forecast errors underlying the entries in Tables 3-5. The bootstrap is applied separately for each subperiod and for each forecast horizon, using a block size of 1 for the $h = 0Q$ forecasts, 2 for $h = 1Q$, and 5 for $h = 1Y$.

4.1 Declining volatility

While there are many nuances in the detailed results, some clear patterns emerge. The univariate RMSEs clearly show the reduced volatility of the economy since the early 1980s. For each horizon, the benchmark univariate RMSEs of GDP growth declined by roughly two-thirds across the 1970-84 and 1985-05 samples (Tables 3-4). The reduced volatility continues to be evident for the inflation measures (Tables 3-4). At the shorter horizons, $h = 0Q$ and $h = 1Q$, the benchmark RMSEs fell by roughly half, but at the longer $h = 1Y$ and (unreported) $h = 2Y$ horizons the variability declined nearly two-thirds. The reverse is true for the interest rate forecasts (Table 5). At the shorter horizons the benchmark RMSEs fell by roughly two-thirds but at the longer horizons the variability declined by less than half.

(2005), for the reasons given in the Corradi and Swanson (2006) and West (2006) surveys of the literature on testing for differences in forecast accuracy. The practical impact of the inclusion of nested models in data snoop applications is unclear; the White and Hansen bootstraps may or may not remain reasonably accurate. Second, for the reasons given in Clark and McCracken (2007), the application to real-time data also violates the technical assumptions of White and Hansen. In this regard, too, the impact is unclear. In pairwise applications, Clark and McCracken find that adjustments in testing necessary in principle for real time data are modest. We suspect the same could be true in multiple model comparisons of the sort considered in this paper.

¹²Because the SPA test is based on t -statistics for equal MSE instead of just differences in MSE, the forecast identified as being significantly best by SPA may not be the forecast with the lowest RMSE ratio. For multi-step forecasts, we compute the variance entering the t -test using the Newey and West (1987) estimator with a lag length of $1.5h$, where h denotes the number of forecast periods.

4.2 Declining predictability

Consistent with the results in such studies as Campbell (2007), there are some clear signs of a decline in the predictability of both output and inflation: it has become harder to beat the accuracy of a univariate forecast. For example, at forecast horizons of $h = 1Y$ or less, most methods or models beat the accuracy of the univariate forecast of GDP growth during the 1970-84 period (Tables 3 and 4). In fact, many do so at a level that is statistically significant; at each horizon Hansen's (2005) SPA test identifies a statistically significant best performer. But over the 1985-2005 period, for $h = 0Q$ and $h = 1Q$ forecasts only the BVAR(4)-TVP models are more accurate at short horizons, and that improvement fails to be statistically significant. At the $h = 1Y$ horizon a handful of the methods continue to outperform the benchmark univariate, but very few are statistically significant.

The predictability of inflation has also declined, although less dramatically than for output. For example, in models with GDP growth and GDP inflation (Table 3), the best 1-year ahead forecasts of inflation improve upon the univariate benchmark RMSE by more than 10 percent in the 1970-84 period but only about 5 percent in 1985-05. The evidence of a decline in inflation predictability is perhaps most striking for CPI forecasts at the $h = 0Q$ horizon. In Table 4, most of the models convincingly outperform the univariate benchmark during the 1970-84 period, with statistically significant maximal gains of roughly 20 percent. But in the following period, fewer methods outperform the benchmark, with gains typically about 4 percent.

Predictability of the T-bill rate has not so much declined as it has shifted to a longer horizon. In Table 5 we see that at the $h = 0Q$ horizon far fewer methods outperform the univariate benchmark as we move from the 1970-84 period to the 1985-05 period. However, the decline in relative predictability starts to weaken as the forecast horizon increases. At the $h = 1Q$ horizon some methods continue to beat the benchmark, although with maximal gains of only about 5 percent. But at the $h = 1Y$ horizon, not only do a larger number of methods improve upon the benchmark, they do so with maximal gains that are substantial and statistically significant, at about 12 percent.

4.3 Averaging methods that typically outperform the benchmark

In light of the considerable sampling error inherent in small-sample forecast comparisons, we shouldn't expect to be able to identify a particular forecast model or method that beats the univariate benchmark for every variable, horizon, and sample period. Instead, we might judge a model or method a success if it beats the univariate benchmark most of the time (with some consistency across the 1970-84 and 1985-05 samples) and, when it fails to do so, is not dramatically worse than the univariate benchmark.

With this consideration in mind, the best forecast would appear to come from the pairwise averaging class: the single best forecast is an average of the univariate forecast with the forecast from a VAR(4) with inflation detrending (a VAR(4) in y , $\pi - \pi_{-1}^*$, and $i - \pi_{-1}^*$, motivated by the work of Kozicki and Tinsley (2001, 2002)). More so than any other forecast, the forecast based on an average of the univariate and inflation-detrended VAR(4) projections beats the univariate benchmark a very high percentage of the time and, when it fails to do so, is generally comparable to the univariate forecast. For example, in the case of forecasts of GDP growth and GDP inflation from models in these variables and the T-bill rate (Table 3), this pairwise average's RMSE ratio is less than 1 for all samples and horizons, with the exception of $h = 0Q$ and $h = 1Q$ forecasts of GDP growth for 1985-05, in which cases the RMSE ratio is only slightly above 1. For 1-year ahead forecasts of GDP growth, the RMSE of this average forecast is about 15 percent below the univariate benchmark for 1970-84 and 9 percent below for 1985-05; the corresponding figures for GDP inflation are roughly 3 percent.

While not quite as good as the average of the univariate and inflation-detrended VAR forecasts, some other averages also seem to perform well, consistently beating the accuracy of the univariate benchmark. In particular, two of the other pairwise forecasts — the VAR(4) with univariate and DVAR(4) with univariate averages — are often, although not always, more accurate than the univariate benchmarks. For instance, in forecasts of GDP growth and CPI inflation (Table 4), these pairwise averages' RMSE ratios are less than 1 in 8 of 12 columns, and only slightly to modestly above 1 in the exceptions. The VAR(4)/univariate average tends to have a more consistent advantage in 1985-05 forecasts. In addition, among the inflation forecasts, the three pairwise combinations (univariate with inflation-detrended

VAR(4), VAR(4) and DVAR(4)) are the most consistent out-performers of the univariate benchmark across both the 1970-84 and 1985-05 subsamples.

The rankings in Table 6 confirm that, from a broad perspective, the best forecasts are simple averages. In these rankings, the single best forecast is the average of the forecasts from the univariate and inflation-detrended VAR(4). Across all variables, horizons, and samples, this forecast has an average ranking of 6.4; the next-best forecast, the average of the univariate and VAR(4) forecasts, has an average ranking of 12.0. While the univariate/inflation-detrended VAR(4) average is, in relative terms, especially good for forecasting the T-bill rate (see column 5), this forecast retains its top rank even when interest rate forecasts are dropped from the calculations (column 2). This average forecast also performs relatively well for forecasting both output (column 3 shows it ranks a close second to the BVAR(4) with inflation detrending) and inflation (column 4 shows it ranks first). As to sample stability, the univariate/inflation-detrended VAR(4) average is best in each of the 1970-84 and 1985-05 samples (columns 6-7).

4.4 Averaging methods that sometimes outperform the benchmark

Among other forecasts, it is difficult to identify any methods that might be seen as consistently equaling or materially beating the univariate benchmark. Take, for instance, the simple equally weighted average of all forecasts, applied to a model in GDP growth, GDP inflation, and the T-bill rate (Table 3). This averaging approach is consistent in beating the univariate benchmark in the 1970-84 sample, but in most cases fails to beat the benchmark in the 1985-05 sample. Similarly, in the case of T-bill forecasts from the same model (Table 5), the all-model average loses out to the univariate benchmark for three of the eight combinations of horizon and sample, while the generally best-performing method of averaging the univariate and inflation-detrended VAR(4) forecasts beats the univariate benchmark in all cases.

A number of the other averaging methods perform quite comparably to the simple average — and thus, by extension, fail to consistently equal or beat (materially) the univariate benchmark. Among the broad average forecasts, from the results in Tables 3-5 there seems to be no advantage of a median or trimmed mean forecast over the simple average. Similarly, MSE-

weighted forecasts are comparable to simple average forecasts, in terms of RMSE accuracy.¹³ For example, in the case of 1-year ahead forecasts of GDP growth and GDP inflation for 1985-05, the recursively MSE-weighted forecast's RMSE ratios are .957 (growth) and 1.028 (inflation), compared to the simple average's ratios of, respectively, .962 and 1.036 (Table 3). In 1-year ahead forecasts of CPI inflation (Table 4), the RMSE ratio of the recursively MSE-weighted forecast is .951 for 1970-84 and 1.055 for 1985-05, compared to the simple average forecast's RMSE ratios of .950 and 1.066, respectively.

Using the best-quartile forecast yields mixed results: the best quartile forecasts are sometimes more accurate and other times less accurate than the simple average and univariate forecasts. For example, in Table 4's results for 1-year ahead forecasts of GDP growth, the best quartile forecast based on a 10 year rolling sample has a RMSE ratio of .780 for 1970-84 and 1.017 for 1985-05, compared to the simple average forecast's RMSE ratios of, respectively, .839 and .997. Where the best quartile forecast seems to have a consistent advantage over a simple average is in output forecasts for 1970-84.

The rankings in Table 6 confirm the broad similarity of the above methods — the simple average, MSE-weighted averages, and best quartile forecasts. For example, the simple average forecast has an overall average ranking of 14.5, compared to rankings of 12.0 for the recursive MSE-weighted forecast and 12.6 for the recursive best quartile forecast. By comparison, the best forecast, the univariate/inflation-detrended VAR(4) average, has an overall ranking of 6.4. In a very broad sense, most of the aforementioned average forecasts are better than the univariate benchmarks in that they all have higher rankings than the univariate's average ranking of 17.3 (column 1). Note, however, that most of their advantage comes in the 1970-84 sample; in the later sample, the univariate forecast generally ranks higher. For instance, for 1970-84 output and inflation forecasts, the all-model average has an average accuracy rank of 13.4, compared to the univariate ranking of 21.8 (column 6). But for 1985-05 forecasts, the all-model average has an average accuracy rank of 16.6, compared to the univariate ranking of 13.9 (column 7).

¹³However, in the case of forecasts of the HP output gap, the MSE-weighted averages are consistently slightly better than the simple averages.

4.5 Averaging methods that rarely outperform the benchmark

Many of the other averaging or combination methods are clearly dominated by univariate benchmarks (and, in turn, other average forecasts). OLS combinations or ridge combinations that approximate OLS often fare especially poorly. The OLS–approximating ridge regression combination (the one with $k = .001$) consistently yields poor forecasts. For example, in the case of 1985-05 1-year ahead forecasts of CPI inflation from models with GDP growth (Table 4), the RMSE ratio of the recursively estimated ridge regression with shrinkage parameter of .001 is 1.458. Similarly, the forecasts based on OLS combination regression using the first and second quartile average forecasts — especially those using rolling samples — are generally dominated by other average forecasts.

While using more shrinkage improves the accuracy of forecast combinations estimated with generalized ridge regression, even the combinations based on ridge regression with non-trivial shrinkage are generally less accurate than the univariate benchmarks and simple average forecasts. For example, in 1985-05 forecasts of GDP growth from models using the GDP inflation measure (Table 3), the RMSE ratios of the $k = 1$ recursive ridge regression forecast are all above those of the simple average forecast. While the ridge forecasts are more commonly beaten by the simple average, there are, to be sure, a number of instances (as in the same example, but with a forecast sample of 1970-84) in which ridge forecasts are more accurate. On balance, though, the ridge combinations seem to be inferior to alternatives such as the simple average forecast.

Forecasts based on using factor model methods to obtain a combination are also generally less accurate than alternatives such as the univariate and simple average forecasts. For example, in the case of 1-year ahead forecasts of GDP growth and GDP inflation for 1985-05, the recursively estimated factor combination forecast’s RMSE ratios are 1.021 (growth) and 1.536 (inflation), compared to the simple average’s ratios of, respectively, .962 and 1.036 (Table 3). The same is true for the PLS forecasts: although PLS forecasts are sometimes more accurate than the simple average, they are often worse. In the same example, the recursive PLS forecast’s RMSE ratios are 1.108 and 1.011, respectively.

The BMA forecasts are also generally, although not universally, dominated by the simple

average. For example, in Table 5’s forecasts of the T-bill rate, the RMSE ratios of the BMA: BIC forecast are consistently above the ratios of the simple average forecast. However, in Table 3’s results for GDP growth and GDP inflation, the accuracy of the BMA: BIC forecast is generally comparable to that of the simple average forecast. Among the alternative BMA forecasts, there are times when those using AIC or PIC to measure the marginal likelihood are more accurate than those using BIC. But more typically, the BMA: BIC forecast is more accurate than the BMA: AIC and BMA: PIC forecasts — the pattern is especially clear in 1985-05 forecasts.

The rankings in Table 6 provide a clear and convenient listing of the forecast methods that are generally dominated by the univariate benchmark and alternatives such as the best-performing pairwise average forecast and the all-model simple average. As previously mentioned, generalized ridge forecasts with little shrinkage ($k = .001$, so as to approximate OLS-based combination) typically perform among the worst forecasts for all horizons, variables and periods, with average ranks consistently in the low- to mid-30s. OLS combinations of quartile forecasts also fare quite poorly when based on rolling samples, with ranks generally in the mid-20s to low 30s. The factor-based combination forecasts are also consistently ranked in the bottom tier, with average rankings generally in the mid-20s. While not necessarily in the bottom tier, the BMA forecasts are generally dominated by the simple average forecast. The overall rankings of the BMA: BIC, BMA: PIC, and BMA: AIC forecasts are 22.0, 25.4, and 29.0, respectively, compared with the simple average forecast’s ranking of 14.5 (first column). The average ranks of the PLS forecasts are consistently around 20 (or much worse in the 5 year rolling case). The ridge-based combination forecasts with the highest degree of shrinkage ($k = 1$) fare much better than the OLS-approximating ridge combinations, but consistently rank below the simple average forecast. For example, as shown in the first column, the 10-year rolling ridge regression with $k = 1$ has an average ranking of 16.9.

4.6 Single VAR methods

Among the single VAR forecasts included for comparison, the BVAR(4) with inflation detrending is generally best. While shrinkage in the form of averaging forecasts from an

inflation-detrended VAR(4) with univariate forecasts is better than estimating the inflation-detrended VAR(4) by Bayesian methods, the latter at least performs comparably to the simple average forecast. For example, as shown in Table 3, forecasts of GDP growth from the BVAR(4) with inflation detrending are often at least as accurate as the simple average forecasts (as, for example, with 1-year ahead forecasts for 1985-05). However, forecasts of GDP inflation from the same model are generally less accurate than the simple average (see, for example, the 1-year ahead forecasts for 1985-05). These examples reflect a pattern evident throughout Tables 3-4: while inflation detrending might be expected to most improve inflation forecasts, it instead most improves output forecasts. Although the accuracy of the other individual VAR models is more variable, overall these models are more clearly dominated by the univariate benchmark and others such as the simple average forecast. For example, in the case of the BVAR(4) using GDP growth and GDP inflation (and the T-bill rate), the simple average forecasts are generally more accurate than the BVAR(4) forecasts of growth over 1970-84, inflation over 1970-84, and inflation over 1985-05 (Table 3).

Consistent with these examples, forecasts from single models are generally dominated by average forecasts. The pattern is clearly evident in the average rankings of Table 6. Across all variables, horizons, and samples, the best-ranked single model is the BVAR(4) with inflation detrending, which is out-ranked by 4 different average forecasts. The other single models rank well below the BVAR(4) with inflation detrending.

While averages are broadly more accurate than single model forecasts, it is less clear that they are consistently more accurate across sample periods. To check consistency, we calculated the correlation of the ranks of all 32 average forecasts and all 50 single model forecasts across the 1970-84 and 1985-05 periods, based on the inflation and output results covered in columns 6-7 of Table 6 (using rankings including T-bill rate forecasts yields essentially the same correlations). The correlation of single model forecast rankings is 53 percent; the correlation of the average forecast rankings is 92 percent. The implication is that not only is the typical average forecast more accurate than the typical single model forecast, it is also consistently so across the two periods.

4.7 Interpretation

Why might simple averages in general and the pairwise average of univariate and inflation-detrended VAR(4) forecasts be more accurate than any single model? As noted in the introduction, in practice it is very difficult to know the form of structural instability, and competing models will differ in their sensitivity to structural change. In such an environment, averages across models are likely to be superior to any single forecast. In line with prior research on combining a range of forecasts that incorporate information from different variables (such as Stock and Watson (1999, 2004)), simple equally weighted averages are typically at least as good as averages based on weights tied to historical forecast accuracy. The limitations of weighted averages relative to simple averages are commonly attributed to difficulties in estimating potentially optimal weights in finite samples, especially when the cross-section dimension is large relative to the time dimension.

As to the particular success of forecasts using inflation detrending, one interpretation is that removing a smooth inflation trend — a trend that matches up well with long-term inflation expectations — from both inflation and the interest rate does a reasonable job of capturing non-stationarities in inflation and interest rates. Kozicki and Tinsley (2001, 2002) have developed such VARs from models with learning about an unobserved, time-varying inflation target of the central bank. Similarly, recent research for other countries has found that related Bayesian methods for centering forecasts around an explicit inflation target (see, e.g., Adolfson, et al. (2007)) improves forecast accuracy.

However, such a single representation is surely not the true model, and noise in estimating the many parameters of the model likely have an adverse effect on forecast accuracy. Therefore, a better forecast can be obtained by applying some form of shrinkage. One approach, which primarily addresses parameter estimation noise, is to use Bayesian shrinkage in estimating the VAR with inflation detrending. Another approach is to combine forecasts from the inflation-detrended VAR with forecasts from an alternative model — in our case, the univariate benchmark (note that the IMA(1) benchmarks for inflation and the T-bill rate imply random walk trends). Koop and Potter (2004) note that such model averaging can be viewed as a form of shrinkage for addressing both parameter estimation noise and model

uncertainty. The superiority of this average forecast can be interpreted as highlighting the value of inflation detrending, shrinkage of parameter noise, and shrinkage to deal with model uncertainty.

5 Conclusion

In this paper we consider a wide range of approaches to averaging VAR forecasts obtained with a variety of primitive methods for managing model instability. Our results indicate that some forms of model averaging consistently improve forecast accuracy. The simplest forms of model averaging — such as those that use equal weights across all models — consistently perform among the best methods.

For forecasting U.S. aggregates today, our results suggest a practical forecaster should put considerable weight on univariate forecasts and pay close attention to trends or low-frequency movements in inflation and interest rates. Our best forecast does both: it simply averages projections from a univariate model and a VAR with inflation and the interest rate centered around an inflation trend that approximates long-run inflation expectations.

References

- Adolfson M, Andersson MK, Linde J, Villani M, Vredin A. 2007. Modern forecasting models in action: improving macroeconomic analyses at central banks. *International Journal of Central Banking*, forthcoming.
- Aiolfi M, Timmermann A. 2006. Persistence in forecasting performance and conditional combination Strategies. *Journal of Econometrics* **135**: 31-54.
- Bates JM, Granger CWJ. 1969. The combination of forecasts. *Operations Research Quarterly* **20**: 451-468.
- Boivin J. 2006. Has U.S. monetary policy changed? Evidence from drifting coefficients and real-time data. *Journal of Money, Credit and Banking* **38**: 1149-1173.
- Campbell SD. 2007. Macroeconomic volatility, predictability, and uncertainty in the Great Moderation: evidence from the Survey of Professional Forecasters. *Journal of Business and Economic Statistics* **25**: 191-200.
- Clark TE, McCracken, MW. 2005a. Combining forecasts from nested models. Manuscript, Federal Reserve Bank of Kansas City.
- Clark TE, McCracken, MW. 2008. Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, forthcoming.
- Clark TE, McCracken, MW. 2006. Forecasting with small macroeconomic VARs in the presence of instability. In *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Rapach DE, Wohar ME (eds). Elsevier: Amsterdam (forthcoming).
- Clark TE, McCracken, MW. 2007. Tests of equal predictive ability with real-time data. Manuscript, Federal Reserve Bank of Kansas City.
- Clements MP, Hendry DF. 1996. Intercept corrections and structural change. *Journal of Applied Econometrics* **11**: 475-494.
- Cogley T. 2002. A simple adaptive measure of core inflation. *Journal of Money, Credit, and Banking* **34**: 94-113.
- Cogley T, Sargent TJ. 2005. Drifts and volatilities: monetary policies and outcomes in the post World War II U.S. *Review of Economic Dynamics* **8**: 262-302.
- Corradi V, Swanson NR. 2006. Predictive density evaluation. In *Handbook of Economic Forecasting*, Elliott G, Granger CWJ, Timmermann A (eds). North Holland.
- Del Negro M, Schorfheide, F. 2004. Priors from general equilibrium models for VARs. *International Economic Review* **45**: 643-673.
- Eklund J, Karlsson S. 2007. Forecast combination and model averaging using predictive measures. *Econometric Reviews* **26**: 329-363.
- Favero C, Marcellino M. 2005. Modelling and forecasting fiscal variables for the Euro Area. *Oxford Bulletin of Economics and Statistics* **67**: 755-783.

- Garratt A, Koop G, Vahey SP. 2006. Forecasting substantial data revisions in the presence of model uncertainty. Reserve Bank of New Zealand Discussion Paper 2006/02.
- Hallman JJ, Porter RD, Small DH. 1991. Is the price level tied to the M2 monetary aggregate in the long run? *American Economic Review* **81**: 841-858.
- Hansen PR. 2005. A test for superior predictive ability. *Journal of Business and Economic Statistics* **23**: 365-380.
- Jacobson T, Jansson P, Vredin A, Warne A. 2001. Monetary policy analysis and inflation targeting in a small open economy: a VAR approach. *Journal of Applied Econometrics* **16**: 487-520.
- Kapetanios G, Labhard V, Price S. 2007. Forecasting using Bayesian and information theoretic model averaging: an application to UK inflation. *Journal of Business and Economic Statistics* **26**: 33-41.
- Koop G, Potter S. 2004. Forecasting in dynamic factor models using Bayesian model averaging. *Econometrics Journal* **7**: 550-565.
- Kozicki S, Hoffman B. 2004. Rounding error: a distorting influence on index data. *Journal of Money, Credit, and Banking* **36**: 319-38.
- Kozicki S, Tinsley PA. 2001. Shifting endpoints in the term structure of interest rates. *Journal of Monetary Economics* **47**: 613-652.
- Kozicki S, Tinsley PA. 2002. Alternative sources of the lag dynamics of inflation. In *Price Adjustment and Monetary Policy*, Bank of Canada Conference Proceedings.
- Litterman RB. 1986. Forecasting with Bayesian vector autoregressions — five years of experience. *Journal of Business and Economic Statistics* **4**: 25-38.
- Newey WK, West KD. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**: 703-708.
- Pesaran MH, Timmermann A. 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics* **137**: 134-161
- Phillips PCB. 1996. Econometric model determination. *Econometrica* **64**: 763-812.
- Robertson J, Tallman E. 2001. Improving federal-funds rate forecasts in VAR models used for policy analysis. *Journal of Business and Economic Statistics* **19**: 324-330.
- Romer CD, Romer DH. 2000. Federal Reserve information and the behavior of interest rates. *American Economic Review* **90**: 429-457.
- Stock JH, Watson MW. 1999. A dynamic factor model framework for forecast combination. *Spanish Economic Review* **1**: 91-121.
- Stock JH, Watson MW. 2004. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* **23**: 405-430.
- Stock JH, Watson MW. 2007. Why has U.S. inflation become harder to forecast? *Journal of*

Money, Credit, and Banking **39**: 3-33.

Timmermann A. 2006. Forecast combinations. In *Handbook of Economic Forecasting*, Elliott G, Granger CWJ, Timmermann A (eds). North Holland.

West KD. 2006. Forecast evaluation. In *Handbook of Economic Forecasting*, Elliott G, Granger CWJ, Timmermann A (eds). North Holland.

White H. 2000. A reality check for data snooping. *Econometrica* **68**: 1097-1126.