Preliminary and incomplete

Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts

Todd E. Clark and Michael W. McCracken*
Federal Reserve Bank of Kansas City and University of Missouri-Columbia

July 2004

Abstract

This paper presents analytical, Monte Carlo, and empirical evidence on the effectiveness of combining recursive and rolling forecasts, compared to using either just a recursive or rolling forecast, when linear predictive models are subject to structural change. We first provide a characterization of the bias-variance tradeoff faced when choosing between either the recursive and rolling schemes or a scalar convex combination of the two. From that, we derive pointwise optimal time-varying and data dependent observation windows and combining weights designed to minimize mean square forecast error. We then proceed to consider other methods of forecast combination including Bayesian methods that shrink the rolling forecast to the recursive and Bayesian model averaging. Monte Carlo experiments and several empirical examples indicate that although the recursive scheme is often difficult to beat, when gains can be obtained, some form of shrinkage can often provide improvements in forecast accuracy relative to forecasts made using the recursive scheme or the rolling scheme with a fixed window width.

Keywords: structural breaks, forecasting, model averaging.

JEL Nos.: C53, C12, C52

*Clark: Economic Research Dept., Federal Reserve Bank of Kansas City, 925 Grand Blvd., Kansas City, MO 64198, todd.e.clark@kc.frb.org. *McCracken (corresponding author)*: Dept. of Economics, University of Missouri-Columbia, 118 Professional Building, Columbia, MO 65211, mccrackenm@missouri.edu. We gratefully acknowledge the excellent research assistance of Taisuke Nakata and helpful comments from Ulrich Muller, Jonathan Wright and participants at the following meetings: MEG, Canadian Economic Assocation, SNDE, MEA, and the conference for young researchers on Forecasting in Time Series. 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 or the Federal Reserve System

1. Introduction

In a universe characterized by heterogeneity and structural change, forecasting agents may feel it necessary to estimate model parameters using only a partial window of the available observations. The intuition is clear. If the earliest available data follow a data-generating process unrelated to the present then using such data to estimate model parameters may lead to biased parameter estimates and forecasts. Such biases can accumulate and lead to larger mean square forecast errors than do forecasts constructed using only that data relevant to the present and (hopefully) future data-generating process. Unfortunately, if we reduce the number of observations in order to reduce heterogeneity we simultaneously increase the variance of the parameter estimates. This increase in variance maps into the forecast errors and causes the mean square forecast error to increase. Hence when constructing a forecast there is a balance between using too much or too little data to estimate model parameters.

This tradeoff leads to patterns in the decisions on whether or not to use all available data when constructing forecasts. As noted in Giacomini and White (2003), the finance literature tends to construct forecasts using only a rolling window of the most recent observations. In the macroeconomics literature, it is more common for forecasts to be constructed recursively – using all available data to estimate unknown parameters (e.g. Stock and Watson, 2003a) – with forecasts only sometimes based on rolling windows. Since both macroeconomic and financial series are known to exhibit periods of structural change (Stock and Watson 1996, Paye and Timmermann 2002), one reason for the rolling approach to be used more often in finance than in macroeconomics may simply be that financial series are often substantially longer. This allows individuals who forecast financial series to make choices to reduce the bias component of the bias-variance tradeoff that macroeconomic forecasters often cannot make due to the potentially sharp deterioration in the variance component of the tradeoff.

In light of the bias-variance tradeoff associated with the choice between a rolling and recursive forecasting scheme, a combination of recursive and rolling forecasts could be superior to the individual forecasts. Indeed, combining recursive and rolling forecasts could be interpreted as yet another form of shrinkage estimation that might be useful in the presence of instabilities. The findings of Min and Zellner (1993), Koop and Potter (2003), Stock and Watson (2003), and Wright (2003) suggest shrinkage (the particular forms vary across these studies) to be effective in samples with instabilities.

Accordingly, we present analytical, Monte Carlo, and empirical evidence on the effectiveness of combining recursive and rolling forecasts, compared to using either just a recursive or rolling forecast. We first provide a characterization of the bias-variance tradeoff involved in choosing between either the recursive and rolling schemes or a scalar convex combination of the two. This tradeoff permits us to derive not only optimal observation windows for the rolling scheme but, given a rolling observation window, the optimal scalar weights for combining the recursive and rolling schemes.

Because we find that simple scalar methods of combining the recursive and rolling forecasts are useful, we also consider combining methods that do not fit directly into our analytical framework. One approach uses standard Bayesian methods to shrink parameter estimates based on a rolling sample toward those based on the recursive sample. Another method consists of using the Bayesian model averaging approach of Wright (2003) to average a recursive forecast with a sequence of rolling forecasts, each with a distinct observation window.

The results in the paper suggest a benefit to some form of combination of recursive and rolling forecasts. Our Monte Carlo and empirical results show that shrinking coefficient estimates based on a rolling window of data seems to be effective and valuable. On average, the shrinkage produces a forecast MSE essentially the same as the recursive MSE when the recursive MSE is best. When there are model instabilities, the shrinkage produces a forecast MSE that often captures most of the gain that can be achieved with the methods we consider. Thus combining recursive and rolling forecasts yields forecasts that are likely to be as good as or better than either recursive or rolling forecasts based on an arbitrary, fixed window size.

Our results build on several lines of extant work. First, we build on the very large literature on forecast combination, a subject that has recently seen renewed interest, both theoretical (e.g. Elliott and Timmermann, 2003) and empirical (e.g. Stock and Watson, 2003a,b). Second, our analysis follows very much in the spirit of Min and Zellner (1993), who also consider forecast combination as a means of handling heterogeneity induced by structural change. Using a Bayesian framework, they combine a stable linear regression model with another with classical unit-root time variation in the parameters.

Finally, our work on the optimal choice of observation window builds on Pesaran and Timmermann (2002). They, too, consider the determinants of the optimal choice of the observation window in a linear regression framework subject to structural change. Using both conditional and unconditional mean square errors as objective functions they find that the optimal length of the observation window is weakly decreasing in the magnitude of the break, weakly decreasing in the magnitude of any change in the residual variance and is weakly increasing in the magnitude of the time since the break date. They derive a recursive data-based stopping rule for selecting the observation window that does not admit a closed-form solution.

We are able to generalize Pesaran and Timmermann's results in many respects – among them, imposing less restrictive assumptions, such as a scalar parameter vector, and obtaining closed form solutions for the optimal window size.

Our paper proceeds as follows. In section 2 we analytically characterize the bias-variance tradeoff and, in light of that tradeoff, determine the optimal observation window. Section 3 details the recursive-rolling combination methods considered. In section 4 we present Monte Carlo evidence on the finite sample effectiveness of our combination approaches and various other methods for forecasting in the potential presence of model instability. Section 5 compares the effectiveness of the various forecast methods approaches in a range of empirical applications. The final section concludes. Various details pertaining to theory and data are presented in Appendixes 1 and 2.

2. Analytical Results on the Bias-Variance Tradeoff and Optimal Observation Window

In this section, after first detailing the necessary notation, we provide an analytical characterization of the bias-variance tradeoff, created by model instability, involved in choosing between recursive and rolling forecasts. In light of that tradeoff, we then derive the optimal rolling observation window. A detailed set of technical assumptions, sufficient for the results, are given in Appendix 1. The same appendix provides general theoretical results (allowing for the recursive and rolling forecasts to be combined with weights α_t and 1- α_t respectively) from which the results in this section are derived as a special case (with $\alpha_t = 0$). We take up the possibility of combining the recursive and rolling forecasts in section 3.

2.1 Environment

The possibility of structural change is modeled using a sequence of linear DGPs of the form

$$y_{T,t+\tau} = x'_{T,t}\beta^*_{T,t} + u_{t+\tau,T} \qquad \beta^*_{T,t} = \beta^* + T^{-1/2}g(t/T)$$

$$Ex_{T,t}u_{T,t+\tau} \equiv Eh_{T,t+\tau} = 0 \text{ for all } t = 1,\ldots,T,\ldots,T+P.$$

Note that the dependent variable $y_{T,t+\tau}$, the predictors $x_{T,t}$ and the error term $u_{T,t+\tau}$ all depend upon T, the initial forecasting origin. This is certainly necessary for the dependent variable since it depends upon the sequence of parameter vectors $\beta^*_{T,t}$ which in turn depend upon T. We allow the predictors and the errors to depend upon T because we do not want to make the strong assumption that these terms are strongly exogenous to the process generating the dependent variable. By doing so we allow the time variation in the parameters to have some influence on their marginal distributions. This is necessary if we want to allow lagged dependent variables to be predictors.

Except where necessary, however, for the remainder we omit the subscript T that is associated with the observables and the errors. Ultimately, the dependence of the observable variables and their marginal distributions on T is not particularly important when the sample size is large. With linear models subject to local structural change and test statistics formed using sample averages of quadratics of the predictors and forecast errors, we are able to take advantage of the concept of asymptotic mean square stationarity as discussed in Hansen (2000). In Appendix 1 we reintroduce the dependence of the observables upon T and more formally keep track of its impact on the bias-variance tradeoff.

At each origin of forecasting t = T,…, T+P, we observe the sequence $\{y_j, x'_j\}_{j=1}^{t}$. These include a scalar random variable y_t to be predicted and a $(k\times1)$ vector of potential predictors x_t which may include lagged dependent variables. Forecasts of the scalar $y_{t+\tau,}$ $t = T,\ldots,T+P$ $\tau \geq 1$, are generated using the vector of covariates x_t and the linear parametric model $x_t'\beta$. The parameters are estimated one of two ways. For a time varying observation window R_t, the parameter estimates satisfy $\hat{\beta}_{R,t}$ = argmin $t^{-1}\sum_{s=1}^{t-\tau}(y_{s+\tau}-x_s'\beta)^2$ and $\hat{\beta}_{L,t}$ = argmin $R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}(y_{s+\tau}-x_s'\beta)^2$ for the recursive and rolling schemes respectively. The corresponding loss associated with the forecast errors are $\hat{u}_{R,t+\tau}^2 = (y_{t+\tau}-x_t'\hat{\beta}_{R,t})^2$ and $\hat{u}_{L,t+\tau}^2 = (y_{t+\tau}-x_t'\hat{\beta}_{L,t})^2$.

Before presenting the results it is useful to provide a brief discussion of Assumptions 1–4 in Appendix 1. In Assumptions 1–3 we maintain that the OLS-estimated DGP is a linear regression subject to local structural change. The local structural change is nonstochastic, square integrable and of a small enough magnitude that the observables are asymptotically mean square stationary. In order to insure that certain weighted partial sums converge weakly to standard Brownian motion W(.), we impose the high level assumption that, in particular, $h_{t+\tau}$ satisfies Theorem 3.2 of De Jong and Davidson (2000). By doing so we also are able to take advantage of various results pertaining to convergence in distribution to stochastic integrals that are also contained in De Jong and Davidson.

Our final assumption is unique. In part (a) of Assumption 4 we generalize assumptions made in West (1996) that require $\lim_{T\to\infty}R_t/T = \lambda_R \in (0, 1)$. Such an assumption is too stringent for our goals. Instead, in parts (a) and (c) we weaken that type of assumption so that $R_t/T \Rightarrow \lambda_R(s) \in (0, s]$, $1 \leq s \leq 1 + \lambda_P$, where $\lim_{T\to\infty}P/T = \lambda_P \in (0, \infty)$ and hence the duration of forecasting is finite but non-trivial. By doing so we permit an observation window that changes with time as evidence of instability is discovered. For the moment we omit a discussion of part (b) but return to it in section 3 when we consider combining the recursive and rolling schemes.

2.2 Theoretical results on the tradeoff: the general case

Our approach to understanding the bias-variance tradeoff is based upon an analysis of $\sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau}-\hat{u}^2_{L,t+\tau})$, the difference in the (normalized) MSEs of the recursive and rolling forecasts.[1] As detailed in Theorem 1 in Appendix 1, we show that this statistic has an asymptotic distribution that can be decomposed into three terms:

$$\sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau}-\hat{u}^2_{L,t+\tau}) \to_d \int_1^{1+\lambda_P}\xi_W(s) = \int_1^{1+\lambda_P}\xi_{W1}(s) + \int_1^{1+\lambda_P}\xi_{W2}(s) + \int_1^{1+\lambda_P}\xi_{W3}(s). \qquad (1)$$

The first component can be interpreted as the pure "variance" contribution to the distribution of the difference in the recursive and rolling MSEs. The third term can be interpreted as the pure "bias" contribution, while the second is an interaction term.

This very general result implies that the bias-variance tradeoff depends on: (1) the rolling window size ($\lambda_R(s)$), (2) the duration of forecasting (λ_P), (3) the dimension of the parameter vector (through the dimension of W or g), (4) the magnitude of the parameter variability (as measured by the integral of quadratics of g), (5) the forecast horizon (as measured by the long-run variance of $h_{t+\tau}$, V) and (6) the second moments of the predictors ($B = \lim_{T\to\infty}(Ex_{T,t}x'_{T,t})^{-1}$).

Providing a more detailed analysis of the distribution of the relative accuracy measure is difficult because we do not have a closed form solution for the density and the bias term allows for very general breaking processes. Therefore, we proceed in the remainder of this section to focus on the mean (rather than the distribution) of the bias-variance tradeoff when there are either no breaks or a single break.

[1] In Theorem 1, the tradeoff is based on $\sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau}-\hat{u}^2_{W,t+\tau})$, which depends upon the combining weights α_t. If we set $\alpha_t = 0$ we find that $\sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau}-\hat{u}^2_{W,t+\tau}) = \sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau}-\hat{u}^2_{L,t+\tau})$.

2.3 The case of no break

We can precisely characterize the mean in the case of no breaks. When there are no breaks we need only analyze the mean of the variance contribution $\int_1^{1+\lambda_P} \xi_{W1}(s)$. Taking expectations and noting that the first of the variance components is zero mean we obtain

$$\int_1^{1+\lambda_P} E\xi_{W1}(s) = \text{tr}(BV)\int_1^{1+\lambda_P} (\frac{1}{s} - \frac{1}{\lambda_R(s)})ds \qquad (2)$$

where tr(.) denotes the trace operator. It is straightforward to establish that all else constant, the mean variance contribution is increasing in the window width $\lambda_R(s)$, decreasing in the forecast duration λ_P and negative semi-definite for all λ_P and $\lambda_R(s)$. Not surprisingly, we obtain the intuitive result that in the absence of any structural breaks the optimal observation window is $\lambda_R(s) = s$. In other words, in the absence of a break it is always best to use the recursive scheme.

2.4 The case of a single break

Suppose that a permanent local structural change, of magnitude $T^{-1/2}g(t/T) = T^{-1/2}\Delta\beta$, in the parameter vector β occurs at time $1 \leq T_B \leq t$ where again, $t = T, \ldots, T+P$ denotes the present forecasting origin. In the following let $\lim_{T\to\infty} T_B/T = \lambda_B \in (0, s)$. Substitution into Theorem 1 in Appendix 1 yields the following corollary regarding the bias-variance tradeoff.

Corollary 2.1: (a) If $\lambda_R(s) > s - \lambda_B$ for all $s \in [1,1+\lambda_P]$ then

$$\int_1^{1+\lambda_P} E\xi_W(s) \ = \ \text{tr}(BV)\int_1^{1+\lambda_P} (\frac{1}{s} - \frac{1}{\lambda_R(s)})ds$$

$$+ \ \Delta\beta' B^{-1}\Delta\beta\int_1^{1+\lambda_P} (s-\lambda_R(s))(s-\lambda_B)(\frac{-(s-\lambda_B)(s+\lambda_R(s))+2s\lambda_R(s)}{s^2\lambda_R^2(s)})ds\,.$$

(b) If $\lambda_R(s) \leq s - \lambda_B$ for all $s \in [1,1+\lambda_P]$ then

$$\int_1^{1+\lambda_P} E\xi_W(s) \;=\; tr(BV)\int_1^{1+\lambda_P}(\frac{1}{s}-\frac{1}{\lambda_R(s)})ds \;+\; \Delta\beta' B^{-1}\Delta\beta\int_1^{1+\lambda_P}(\frac{\lambda_B^2}{s^2})ds\,.$$

From Corollary 2.1 we see that the tradeoff depends upon a weighted average of the precision of the parameter estimates as measured by tr(BV) and the magnitude of the structural break as measured by the quadratic $\Delta\beta' B^{-1}\Delta\beta$. Note that the first term in each of the expansions is negative semi-definite while that for the latter is positive semi-definite. Intuitively, we would expect this to imply that as the magnitude of the break increases relative to the precision of the parameter estimates it is optimal to decrease the observation window.

Even so, part (b) of Corollary 2.1 places a bound on the amount that one would be willing to decrease the width of the observation window. This can be seen if we notice that the bias component does not depend upon the magnitude of the observation window – except for the condition that it be less than or equal to $s - \lambda_B$. Since the variance component is monotone increasing in $\lambda_R(s)$ we immediately conclude that it is never optimal to choose an observation window less than the entirety of the sample since the most recent break.

Although it is perhaps not immediately obvious, Corollary 2.1 does not consider all possible choices of $\lambda_R(s)$. There are three possibilities. In the first the window always is large enough to include the break date. In the second the window is never big enough to contain observations prior to the break date. In the third, the window is allowed to sometimes capture the break date but sometimes not. We do not consider the third case since it is irrelevant towards our goal of optimizing over the bias-variance tradeoff. This follows from our intuitive discussion of part (b) of Corollary 2.1 and is stated explicitly in the following corollary.

Corollary 2.2: In the presence of a single break in the regression parameter vector, the pointwise optimal observation window satisfies

$$\lambda_R^*(s) = \begin{cases} s & \dfrac{s}{2\lambda_B(s-\lambda_B)} \geq \dfrac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)} \\ \dfrac{2(s-\lambda_B)^2}{2(s-\lambda_B)-(\dfrac{tr(BV)}{\Delta\beta' B^{-1}\Delta\beta})} & \dfrac{s}{2\lambda_B(s-\lambda_B)} < \dfrac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)} \\ s-\lambda_B & s-\lambda_B \rightarrow \infty \end{cases} .$$

Corollary 2.2 provides pointwise optimal observation windows for forecasting in the presence of a single structural change in the regression coefficients. We describe these as pointwise optimal because they are derived by maximizing the individual elements from part (a) of Corollary 2.1 that contribute to the average expected mean square differential over the duration of forecasting. In other words, the results of Corollary 2.2 follow from maximizing

$$tr(BV)(\frac{1}{s}-\frac{1}{\lambda_R(s)}) + \Delta\beta' B^{-1}\Delta\beta(s-\lambda_R(s))(s-\lambda_B)(\frac{-(s-\lambda_B)(s+\lambda_R(s))+2s\lambda_R(s))}{s^2\lambda_R^2(s)}) \qquad (3)$$

with respect to $\lambda_R(s)$ for each s. Again, we do not need to consider the case in which $\lambda_R(s)$ is *ever* less than s – λ_B since, as can be seen from part (b) of Corollary 2.1, any time period in which $\lambda_R(s) < s - \lambda_B$ monotonically reduces the objective function.

The formula in Corollary 2.2 is plain enough that comparative statics are reasonably simple. Perhaps the most important is that the observation window is increasing in the ratio $tr(BV)/\Delta\beta' B^{-1}\Delta\beta$. For smaller breaks we expect to use a larger observation window and when parameter estimates are more precisely estimated (so that tr(BV) is small) we expect to use a smaller observation window.

Note, however, that the term $\Delta\beta^{'}B^{-1}\Delta\beta$ is a function of the local break magnitudes $\Delta\beta$ and not the global break magnitudes we estimate in practice. Moreover, note that these optimal windows are not presented relative to an environment in which agents are forecasting in 'real time'. Taken literally, if we were to use the formulas we would need to know the original forecasting origin T and estimate quantities like s and λ_B using $\hat{s} = t/T$ and $\hat{\lambda}_B = \hat{T}_B/T$ for an estimated break date $\hat{T}_B$.

Rather than take this approach we suggest a transformed formula that treats the present forecasting origin as the only one that is relevant. Let $\hat{B}$ and $\hat{V}$ denote estimates of B and V respectively. If for an estimated *global* break $\Delta\hat{\beta}$ at an estimated break date $\hat{T}_B$, we let $\Delta\hat{\beta}$ denote an estimate of the *local* change in β $(\Delta\beta/T^{1/2})$ at time T_B and $\hat{\delta}_B = \hat{T}_B / t$, we obtain the following real time estimate of the pointwise optimal observation window.[2]

$$\hat{R}_t^* = \begin{cases} t & \dfrac{1}{2\hat{\delta}_B(1-\hat{\delta}_B)} \geq \dfrac{t\Delta\hat{\beta}^{'}\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})} \\ \dfrac{2t(1-\hat{\delta}_B)^2}{2(1-\hat{\delta}_B)-(\dfrac{tr(\hat{B}\hat{V})}{t\Delta\hat{\beta}^{'}\hat{B}^{-1}\Delta\hat{\beta}})} & \dfrac{1}{2\hat{\delta}_B(1-\hat{\delta}_B)} < \dfrac{t\Delta\hat{\beta}^{'}\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})} \\ t(1-\hat{\delta}_B) & 1-\hat{\delta}_B \rightarrow 1 \end{cases} . \tag{4}$$

[2] We estimate B with $\hat{B}=(t^{-1}\sum_{j=1}^{t} x_j x_j^{'})^{-1}$, where x_t is the vector of regressors in the forecasting model (supposing the MSE stationarity assumed in the theoretical analysis. In the Monte Carlo experiments, tr(BV)is estimated imposing homoskedasticity: $tr(BV) = k\hat{\sigma}^2$, where k is the number of regressors in the forecasting model and $\hat{\sigma}^2$ is the estimated residual variance of the forecasting model estimated with data from 1 to t. In the empirical applications, though, we use the estimate $tr(BV) = tr[(t^{-1}\sum_{j=1}^{t} x_j x_j^{'})^{-1}(t^{-1}\sum_{j=1}^{t} \hat{u}_{j+\tau}^2 x_j x_j^{'})]$, where $\hat{u}$ refers to the residuals from estimates of the forecasting model using data from 1 to t.

One final note on the formulae in Corollary 2.2 and (4). In Corollary 2.2, we use local breaks to model the bias-variance tradeoff faced by a forecasting agent in finite samples. By doing so we are able to derive closed form solutions for the optimal observation window. Unfortunately, by taking this approach we have created a situation where we have no choice but to use inconsistent (Bai (1997)) global break magnitudes and dates to estimate the assumed local magnitudes. Not surprisingly, in our Monte Carlo results we find that experiments which treat the break magnitudes and dates as known perform better than other experiments where we estimate these quantities. Even so, we find that the estimated quantities perform well enough to be a valuable tool for forecasting.

3. Approaches to Combining Recursive and Rolling Forecasts

In section 2 we discussed how the choice of observation window can improve forecast accuracy by appropriately balancing a bias-variance tradeoff. In this section, we consider whether combining recursive and rolling forecasts can also improve forecast accuracy by balancing a similar tradeoff. We do so using three different combination approaches. The first is a simple scalar combination of recursive and rolling forecasts. The second, which can be viewed as a matrix-valued combination, is based on Bayesian shrinkage of rolling model estimates toward recursive estimates. The third relies on Bayesian model averaging, as implemented in Wright (2003).

3.1 Simple scalar combination

The simplest possible approach to combination is to form a scalar linear combination of recursive and rolling forecasts. With linear models, of course, the linear combination of the forecasts is the same as that generated with a linear combination of the recursive and rolling parameter estimates. Accordingly, we consider generating a forecast using coefficients $\hat{\beta}_{W,t}$ = $\alpha_t\hat{\beta}_{R,t} + (1-\alpha_t)\hat{\beta}_{L,t}$, with corresponding loss $\hat{u}^2_{W,t+\tau} = (y_{t+\tau}-x'_t\hat{\beta}_{W,t})^2$. The size of the observation window used by the rolling coefficient estimates $\hat{\beta}_{L,t}$ could be set arbitrarily at, say, 40 observations, or it could be chosen to produce the lowest possible MSE of the combined forecast.

Using Theorem 1 in Appendix 1, we are able to derive these optimal combining weights in the presence of a single structural break. Moreover, we are able to generalize the analytical results from Corollary 2.2 on the optimal observation window to an environment in which we choose the optimal observation window for a given combining weight. If, as we have for the observation window R_t, we let α_t converge weakly to the function $\alpha(s)$, the following corollaries provide the desired results. For each we maintain the same assumptions and notation used in Corollaries 2.1 and 2.2.

Corollary 3.1: (a) If $\lambda_R(s) > s - \lambda_B$ for all $s \in (1, 1+\lambda_P]$ then

$$\int_1^{1+\lambda_P} E\xi_W(s) = tr(BV)\int_1^{1+\lambda_P}(1-\alpha(s))^2(\frac{1}{s}-\frac{1}{\lambda_R(s)})ds$$

$$+\ \Delta\beta'B^{-1}\Delta\beta\int_1^{1+\lambda_P}(1-\alpha(s))(s-\lambda_R(s))(s-\lambda_B)(\frac{(s-\lambda_B)(\alpha(s)(s-\lambda_R(s))-(s+\lambda_R(s)))+2s\lambda_R(s)}{s^2\lambda_R^2(s)})ds.$$

(b) If $\lambda_R(s) \leq s - \lambda_B$ for all $s \in (1, 1+\lambda_P]$ then

$$\int_1^{1+\lambda_P} E\xi_W(s) = tr(BV)\int_1^{1+\lambda_P}(1-\alpha(s))^2(\frac{1}{s}-\frac{1}{\lambda_R(s)})ds\ +\ \Delta\beta'B^{-1}\Delta\beta\int_1^{1+\lambda_P}(1-\alpha^2(s))(\frac{\lambda_B^2}{s^2})ds.$$

Corollary 3.2: In the presence of a single break in the regression parameter vector, the pointwise optimal window width and combining weights satisfy

$$\lambda_R^*(s,\alpha(s)) = \begin{cases} s & \dfrac{s(1-\alpha(s))}{2\lambda_B(s-\lambda_B)} \geq \dfrac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)} \\ \dfrac{2s(1-\alpha(s))(s-\lambda_B)^2}{2(s-\lambda_B)(s-\alpha(s)(s-\lambda_B))-s(1-\alpha(s))(\dfrac{tr(BV)}{\Delta\beta' B^{-1}\Delta\beta})} & \dfrac{s(1-\alpha(s))}{2\lambda_B(s-\lambda_B)} < \dfrac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)} \\ s-\lambda_B & s-\lambda_B \to \infty \end{cases}$$

and for $\lambda_R(s) > s-\lambda_B$

$$\alpha^*(s,\lambda_R(s)) = \begin{cases} \dfrac{s\lambda_R(s)+(\dfrac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)})s(s-\lambda_B)(s-\lambda_B-\lambda_R(s))}{s\lambda_R(s)+(\dfrac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)})(s-\lambda_B)^2(s-\lambda_R(s))} & \dfrac{(s-\lambda_B)^2}{(s-\lambda_B)-(\dfrac{tr(BV)}{\Delta\beta' B^{-1}\Delta\beta})} \geq \lambda_R(s) \\ 0 & \dfrac{(s-\lambda_B)^2}{(s-\lambda_B)-(\dfrac{tr(BV)}{\Delta\beta' B^{-1}\Delta\beta})} < \lambda_R(s) \end{cases} .$$

The Corollary provides pointwise optimal observation windows and combining weights for forecasting in the presence of a single structural change in the regression coefficients. As was the case in Corollary 2.2, we describe these as pointwise optimal because they are derived by maximizing the individual elements from part (a) of Corollary 3.1. Again, we do not need to consider the case in which $\lambda_R(s)$ is *ever* less than $s - \lambda_B$ since any time period in which $\lambda_R(s) < s - \lambda_B$ monotonically reduces the objective function.

The formula for the optimal observation window $\lambda_R(s)$, for a given combining weight $\alpha(s)$, is very closely related to that from Corollary 2.2. Perhaps the largest difference between the two is that the observation window on the rolling component is more likely to be smaller since the bias-variance tradeoff is affected by the fact that some weight ($\alpha(s) > 0$) is being placed on the recursive forecast. Comparative statics for the combining weights are also relatively straightforward. As the observation window on the rolling component increases, we place less weight on the recursive scheme. Similarly, as the magnitude of the break increases relative to the precision of the parameter estimates, we also place less weight on the recursive scheme. Finally, we obtain the intuitive result that as the time since the break increases (s–λ_B), we eventually place all weight on the rolling scheme.

As was the case for the formula from Corollary 2.2, the optimal observation windows and combining weights in Corollary 3.2 are not presented in a real time context and depend upon several unknown quantities. If we make the same change of scale and use the same estimators that were used for equation (4), we obtain the real time equivalents of the formulas from Corollary 3.2.

$$\hat{R}_t^*(\alpha_t) = \begin{cases} t & \dfrac{(1\text{-}\alpha_t)}{2\hat{\delta}_B(1\text{-}\hat{\delta}_B)} \geq \dfrac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})} \\ \dfrac{2t(1\text{-}\alpha_t)(1\text{-}\hat{\delta}_B)^2}{2(1\text{-}\hat{\delta}_B)(1\text{-}\alpha_t(1\text{-}\hat{\delta}_B))\text{-}(1\text{-}\alpha_t)(\dfrac{tr(\hat{B}\hat{V})}{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}})} & \dfrac{(1\text{-}\alpha_t)}{2\hat{\delta}_B(1\text{-}\hat{\delta}_B)} < \dfrac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})} \\ t(1\text{-}\hat{\delta}_B) & 1-\hat{\delta}_B \to 1 \end{cases} \tag{5}$$

and

$$\hat{\alpha}_t^*(R_t) = \begin{cases} \dfrac{(R_t/t)+(\dfrac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})})(1\text{-}\hat{\delta}_B)(1\text{-}\hat{\delta}_B\text{-}(R_t/t))}{(R_t/t)+(\dfrac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})})(1\text{-}\hat{\delta}_B)^2(1\text{-}(R_t/t))} & \dfrac{t(1\text{-}\hat{\delta}_B)^2}{(1\text{-}\hat{\delta}_B)\text{-}(\dfrac{tr(\hat{B}\hat{V})}{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}})} \geq R_t \\ 0 & \dfrac{t(1\text{-}\hat{\delta}_B)^2}{(1\text{-}\hat{\delta}_B)\text{-}(\dfrac{tr(\hat{B}\hat{V})}{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}})} < R_t \end{cases} . \tag{6}$$

One final note on Corollary 3.2. The optimal solutions for the observation window λ_R(s) and combining weight α(s) are each derived conditioning on the other. An alternative approach would be to derive the jointly optimal observation and combining weights. We have attempted to do so but jointly determining both optimal values does not appear readily tractable.

3.2 A Bayesian shrinkage forecast

Given the bias-variance tradeoff between recursive and rolling forecasts, a second combination approach that might seem natural is to use parameter estimates based on a rolling sample shrunken so as to reduce the noise in the parameter estimates and resulting forecast. We therefore consider shrinking rolling sample estimates toward the recursive estimates, implemented with standard Bayesian formulae. Recall that for a prior $\beta \sim N(m, \sigma^2 M)$, the Normal linear regression model yields the posterior maximizing estimate $\tilde{\beta}$ = $(M^{-1}+X'X)^{-1}(M^{-1}m+X'Y)$ where X denotes the relevant design matrix and Y the associated vector of dependent variables. If we treat the recursive parameter estimates as the prior mean and treat the associated standard errors under conditional homoskedasticity as our prior variance we have $m = \hat{\beta}_{R,t}$ and $M = B_R^{-1}(t)$ where $B_R(t) = (t^{-1}\sum_{j=1}^{t-\tau} x_j x_j')^{-1}$. If we let $B_L(t) = (R_t^{-1}\sum_{j=t-\tau-R_t+1}^{t-\tau} x_j x_j')^{-1}$, our Bayesian shrinkage estimator then follows by constructing the posterior maximizing rolling parameter estimates given this prior:

$$\begin{aligned}\tilde{\beta}_{W,t} &= [tB_R(t)^{-1} + R_tB_L(t)^{-1}]^{-1}[tB_R(t)^{-1}\hat{\beta}_{R,t} + \textstyle\sum_{s=t-\tau-R_t+1}^{t-\tau} x_s y_{s+\tau}] \\ &= [B_R(t)^{-1} + (R_t/t)B_L(t)^{-1}]^{-1}B_R(t)^{-1}\hat{\beta}_{R,t} + [(t/R_t)B_R(t)^{-1} + B_L(t)^{-1}]^{-1}B_L(t)^{-1}\hat{\beta}_{L,t}. \qquad (7)\end{aligned}$$

It is clear from the right-hand side of (7) that the parameter estimates are a linear combination of both recursive and rolling parameter estimates. In contrast to the simple combination considered in our analytical work, here the weights are matrix valued and depend upon the ratio R_t/t and the matrices of sample second moments $B_R(t)$ and $B_L(t)$.

This Bayesian shrinkage estimator of course involves selecting a rolling observation window. One approach is to use an arbitrary window of, say, 40 observations. Another approach is to exploit the fact that the shrinkage estimator (7) can be fit into our general analytical framework, and use the appropriate optimal window. Under the assumption of asymptotic mean square stationarity, our Bayesian shrinkage estimator is asymptotically equivalent to a scalar-weighted combination of recursive and rolling estimators, with combination weights $\alpha_t = t/(t + R_t)$ and $(1 - \alpha_t) = R_t/(t + R_t)$:

$$\begin{aligned}\tilde{\beta}_{W,t} &= [B_R(t)^{-1} + (R_t/t)B_L(t)^{-1}]^{-1}B_R(t)^{-1}\hat{\beta}_{R,t} + [(t/R_t)B_R(t)^{-1} + B_L(t)^{-1}]^{-1}B_L(t)^{-1}\hat{\beta}_{L,t} \qquad (8)\\ &\sim [1 + (R_t/t)]^{-1}\hat{\beta}_{R,t} + [(t/R_t) + 1]^{-1}\hat{\beta}_{L,t}\\ &= (\frac{t}{t+R_t})\hat{\beta}_{R,t} + (\frac{R_t}{t+R_t})\hat{\beta}_{L,t}\,.\end{aligned}$$

To obtain the asymptotically optimal observation window associated with these weights we need only substitute these weights into our formula from equation (5) . Doing so we find that

$$\hat{R}_t^*(t/(t+R_t)) = \begin{cases} t & \frac{1}{4\hat{\delta}_B(1-\hat{\delta}_B)} \geq \frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})} \\ \frac{2t[(1-\hat{\delta}_B)^2 - \hat{\delta}(1-\hat{\delta}_B)]}{2(1-\hat{\delta}_B)-(\frac{tr(\hat{B}\hat{V})}{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}})} & \frac{1}{4\hat{\delta}_B(1-\hat{\delta}_B)} < \frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})} < \frac{1}{2\hat{\delta}_B}\,. \\ t(1-\hat{\delta}_B) & \frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})} \geq \frac{1}{2\hat{\delta}_B} \end{cases} \qquad (9)$$

The formula in (9) is very close to that in (4) when $\alpha_t = 0$, but with one important difference. We are less likely to have a solution where R_t is close to t and much more likely to have a solution where $R_t = t(1-\hat{\delta}_B)$. Since the Bayesian shrinkage method always places some weight on the recursive component it is not surprising that the observation window for the rolling component of the shrinkage estimator is smaller than that for the pure rolling ($\alpha_t = 0$) scheme.

3.3 Bayesian model averaging

Yet another approach to shrinking rolling forecasts toward the recursive might be to average a recursive forecast with forecasts generated with a potentially wide range of different estimation samples. Bayesian model averaging (BMA) of the form considered by Wright (2003) provides a natural way of doing so. At each forecast date t, suppose that a single, discrete break in the full set of model coefficients could have occurred at any point in the past (subject to some trimming of the start of the sample and the end of the sample, as is usually required in break analysis). In our implementation, allowing for the possibility of a single break point anywhere between observations 20 through t-20 implies a total of t-39 models with a break. For each time t, the forecast generated by a model with a break in all coefficients at date t_B and estimated with all data up to t is of course exactly the same as the forecast generated from a model estimated with just data starting in t_B+1. Therefore, applying BMA techniques to obtain a forecast averaged across the recursive model and the models with breaks (each model represents a different characterization of observations 1 to t) is the same as averaging across the recursive forecast and rolling forecasts based on different observation windows.

In the particulars of our implementation of BMA, we largely follow the settings of Wright (2003). We estimate each forecast model by least squares (which of course can be viewed as Bayesian estimation with a diffuse prior) and use Bayesian methods simply to weight the forecasts. In the benchmark case, the prior probability, Prob(M_i), on each model is just 1/the number of models. We also consider the alternative of putting a large prior on the recursive forecast – a weight of .7 – and a weight of .3/the number of models on each of the rolling forecasts. In calculating the posterior probabilities, Prob(M_i|data), of each model, we set the prior on the coefficients equal to the recursive estimates.[3] Specifically, at each forecast origin t we calculate the posterior probability of each model M_i using

$$\text{Prob}(M_i|\text{data}) \;=\; \frac{\text{Prob}(\text{data}|M_i)\times\text{Prob}(M_i)}{\sum_i \text{Prob}(\text{data}|M_i)\times\text{Prob}(M_i)},$$

where:

$$\text{Prob}(\text{data}|M_i) \;\propto\; (1+\phi)^{-p_i/2} S_i^{-(t+1)}$$

ϕ = parameter determining the rate of shrinkage toward the prior

p_i = the number of explanatory variables in model i

$$S_i^2 \;=\; (Y - Z_i\hat{\Gamma}_i)'(Y - Z_i\hat{\Gamma}_i) + \frac{1}{1+\phi}(\hat{\Gamma}_i - \Gamma_{\text{prior}})'Z_i'Z_i\,(\hat{\Gamma}_i - \Gamma_{\text{prior}})$$

Z_i = matrix of variables in model i (including x_s and, in the models used to generate rolling forecasts, x_s interacted with a break dummy)

$\hat{\Gamma}_i$ = OLS-estimated coefficients of model i

Γ_{prior} = recursive estimates of the coefficients on the x_s variables and zeros for the break terms in the model.

4. Monte Carlo Results

We use Monte Carlo simulations of simple bivariate data-generating processes to evaluate, in finite samples, the performance of the forecast methods described above. In these experiments, the DGP relates the predictand y to lagged y and lagged x, with the coefficients on lagged y and

[3] As Wright (2003) actually uses a coefficient prior of 0, our use of the recursive prior requires a simple adjustment to the S term that enters the posterior probability.

x subject to a structural break. As described below, forecasts of y are generated with the basic approaches considered above, along with some related methods that are used or might be used in practice. Performance is evaluated using some simple summary statistics of the distribution of each forecast's MSE: the average MSE across Monte Carlo draws (medians yield very similar results), along with the probability of equaling or beating the recursive forecast's MSE.

4.1 Experiment design

The DGPs considered share the same basic form, differing only in the persistence of the predictand y and the size of the coefficient break:

$$\begin{aligned} y_t &= (b_y + d_{t-1}\Delta b_y)y_{t-1} + (.5 + d_{t-1}\Delta b_x)x_{t-1} + u_t \\ x_t &= .5x_{t-1} + v_t \\ u_t, v_t &\text{ iid } N(0,1) \\ d_t &= 1(t \geq \lambda_B T). \end{aligned}$$

We begin by considering forecast performance in two stable models, one with $b_y = .3$ (DGP 1-S) and another with $b_y = .9$ (DGP 2-S), imposing $\Delta b_y = \Delta b_x = 0$ in both cases. We then consider four specifications with breaks:

DGP 1-B1	$b_y = .3$	$(\Delta b_y, \Delta b_x) = (-.3, -.5)$
DGP 2-B1	$b_y = .9$	$(\Delta b_y, \Delta b_x) = (-.3, -.5)$
DGP 1-B2	$b_y = .3$	$(\Delta b_y, \Delta b_x) = (0, -.5)$
DGP 2-B2	$b_y = .9$	$(\Delta b_y, \Delta b_x) = (0, -.5)$.

For DGPs with breaks, we present results for experiments with three different break dates (a single break in each experiment): λ_B = .6, .8, and 1.

In each experiment, we conduct 1000 simulations of data sets of 200 observations (not counting the initial observation necessitated by the lag structure of the DGP). The data are generated using innovation draws from the standard normal distribution and the autoregressive structure of the DGP.[4] We set T, the number of observations preceding the first forecast date, to 100, and consider forecast periods of various lengths: $\lambda_P = .2, .4, .6, 1.0$. For each value of λ_P, forecasts are evaluated over the period T+1 through $(1 + \lambda_P)$T.

4.2 Forecast approaches

Forecasts of y_{t+1}, t = T,…, T+P are formed from various estimates of the model

$$y_t = \gamma_0 + \gamma_1 y_{t-1} + \gamma_2 x_{t-1} + e_t \,.$$

The model estimation or forecasting approaches, detailed in Table 1, include rolling estimation using various windows of arbitrary, fixed size, along with this paper's proposed Bayesian shrinkage of the arbitrary, rolling estimates toward the recursive. The approaches also include rolling estimation using the optimal window size R_t^* defined in equation (4) and Bayesian shrinkage using the optimal shrinkage window size R_t^* defined in (9). Although infeasible in empirical applications, the performance of forecasts based on R_t^* calculated using the known features of the DGP – the break point, the break size, and the population moments of the data – provides a useful benchmark of the potential gains to forecasting with a rolling window.[5]

We also consider a range of forecasts based on "real time" estimates of breaks in the forecasting model.[6] The break tests are based on the full set of forecast model coefficients.

[4] The initial observations necessitated by the lag structure of the model are generated from draws of the unconditional normal distribution implied by the (pre-break) model parameterization.

[5] In calculating the "known" R_t*, we set the change in the vector of forecast model coefficients to $\Delta\beta = \sqrt{T}(0 \quad \Delta b_y \quad \Delta b_x)'$ (the local alternative assumed in generating (4) means a finite-sample break needs to be scaled by $\sqrt{T}$) and calculate the appropriate second moments using the population values implied by the pre-break parameterization of the model.

[6] As noted by Inoue and Rossi (2003), repeated application of break tests in such real time analyses with the use of standard critical values will result in spurious break findings. Using the adjusted critical values provided by Inoue

Under these approaches, the forecast at each time t+1 is generated by first testing for a break in data up through period t and then estimating the forecasting model with post-break data through t. For all of the break metrics, we impose a minimum segment length of 20 periods. In general, each of these break-based methods yields a rolling forecast based on a window of time-varying size. But if in forecast period t+1 the break metric fails to identify a break in earlier data, then the estimation window is the full, available sample, and the forecast for t+1 is the same as the recursive forecast. That said, in light of concerns over the power of break tests in small samples, it might be natural to simply use the estimated break and break date without requiring the break to be statistically significant. Although not reported in this paper in the interest of brevity, Monte Carlo results indicate that such an approach can be very effective when the DGP truly has a break, but perform poorly when the DGP is nearly or entirely stable. However, we do find that allowing a break regardless of statistical significance can be effective when combined with Bayesian shrinkage, and we report these results. The method is detailed below.

The first break dating approach we consider is the reverse order CUSUM (of squares) method proposed by Pesaran and Timmermann (2002), which involves searching backward from each forecast date to find the most recent break and then estimating the forecasting model with just post-break data.[7] Because the reverse CUSUM proves to be prone to spurious break findings, a relatively parsimonious 1 percent significance level is used in identifying breaks with the CUSUM of squares.[8] Second, we consider two different forecasts based on single breaks

and Rossi would improve the stable-DGP performance of some of our break test-based methods. But in DGPs with breaks, performance would deteriorate.

[7] For data samples of up to a little more than 200 observations, our CUSUM analysis uses the asymptotic critical values provided by Durbin (1969) and Edgerton and Wells (1994). For larger data samples, our CUSUM results rely on the asymptotic approximation of Edgerton and Wells.

[8] In results not reported in the interest of brevity, we also followed Pesaran and Timmermann (2002) and used the BIC criterion of Yao (1988) and Bai and Perron (2003) to determine the optimal number and dates of potentially multiple breaks and then estimate the forecasting model with data subsequent to the most recent break. We omit the results because they are comparable to those reported for the single break supWald approach.

identified with Andrews' (1993) sup Wald test. One forecast uses a model estimated with just data since a break (dated by its least squares date estimate) identified as significant by the Andrews test at the 5% level.[9] The other forecast is based on estimates from a rolling window of R_t^* observations, where R_t^* is estimated using equation (4). In this case, of course, while the size of the rolling window is determined by the break date and size, by design the window often includes some pre-break data.

We also consider Bayesian shrinkage forecasts based on the rolling windows determined with the sup Wald break metric. Specifically, a forecast is formed using model estimates obtained by Bayesian shrinkage of coefficients estimated with data since the sup-Wald - identified break. In practice, this approach turns out to be the same as a Bayesian shrinkage forecast based on an Andrews test-determined estimate of the optimal shrinkage window size R_t^* defined in (9). As noted above, this optimal window is essentially always just the post-break window, resulting in the equivalence to the approach of shrinking rolling estimates based on an Andrews test-identified window. In light of the potentially low power of the break test, we also construct a Bayesian shrinkage forecast based on an optimal shrinkage window size R_t^* (Bayes α) calculated without regard to the statistical significance of the break.

Finally, we consider BMA forecasts that allow for the possibility of a single break in the vector of model coefficients anytime between observations 20 and t-20, where t denotes the forecast origin. Although we have experimented with various values of the parameter ϕ that determines the rate of shrinkage toward the recursive (a smaller value corresponds to more

[9] At each point in time, the asymptotic p-value of the sup Wald test is calculated using Hansen's (1997) approximation. Because the size of the sample (*t*) increases as forecasting moves forward in time, and the minimum break segment length is held at 20 observations, the value of the sample trim parameter π_0, which the asymptotic distribution depends on, decreases with time. Our break test evaluation takes the time variation in π_0 into account.

shrinkage) used in calculating the posterior probabilities, we report results for the single value that seems to work best: $\phi = .2$.

Note also that, for comparison, we include forecasts based on models estimated with discounted least squares (with a discount rate of .99), widely used in analysis of macroeconomic models with learning.

4.3 Simulation results

In our Monte Carlo comparison of forecast approaches, we mostly base our evaluation on average MSEs over a range of forecast samples. For simplicity, in presenting average MSEs, for only the recursive forecast do we report actual average MSEs. For all other forecasts, we report the ratio of a forecast's average MSE to the recursive forecast's average MSE. To capture potential differences across approaches in MSE distributions, we also present some evidence on the probabilities of equaling or beating a recursive forecast.

4.3.1 Stable DGPs: Average MSEs

With stable DGPs, the most accurate forecasting scheme will of course be the recursive. Moreover, because the DGP has no break, the optimal rolling window R_t^* (α =0) in (4) will be the full sample, so both the rolling forecast based on the known R_t^* (α =0) and the Bayesian shrinkage forecast based on the known optimal Bayesian R_t^* (Bayes α) will be the same as the recursive forecast.

Not surprisingly, then, the average MSEs reported in Table 2 from simulations of the stable DGPs (DGP 1-S and DGP 2-S) show that no forecast beats the recursive forecast – all of the reported MSE ratios are 1.000 or higher. Using an arbitrary rolling window yields considerably less accurate forecasts, with the loss bigger the smaller the window. For example, with DGP 2-S and a forecast sample of 20 observations ($\lambda_P = .2$), using a rolling estimation window of 20 observations yields, on average, a forecast with MSE 20.2 percent larger than the recursive forecast's MSE.

Forecasts with rolling windows determined by formal break tests perform considerably better, with their performance ranking determined by the break metrics' relative parsimony. The reverse CUSUM approach yields a forecast modestly less accurate than a recursive forecast, with the accuracy loss rising as the forecast period gets longer. For example, with DGP 2-S and a forecast sample of 40 observations ($\lambda_P = .4$), the reverse CUSUM forecast has an average MSE 1.7 percent larger than the recursive forecast. For the same DGP and number of forecast periods, a forecast based on the sup Wald break test outcome (*rolling: sup Wald R*) has an average MSE 3.1 percent greater than the recursive forecast. Similarly, the rolling forecast based on an estimate of R_t^* ($\alpha = 0$) is modestly less accurate than the recursive, much more so for the higher persistence DGP 2-S than DGP 1-S. In all, such findings highlight the crucial dependence of these methods on the accuracy of the break metrics.

For all forecasts based on a rolling window of data, using Bayesian shrinkage of the rolling coefficient estimates toward the recursive effectively eliminates any loss in accuracy relative to the recursive forecast. As shown in Table 2, shrinkage of model estimates based on arbitrary rolling windows of 20 or 40 observations yields forecasts with average MSE no worse than .3 percent larger than the recursive forecasts. Shrinkage of model estimates using a sup Wald-determined rolling window yields a forecast (*shrinkage: sup Wald R*) that, at worse, has an average MSE .1 percent larger than the recursive. As indicated by the results in the *shrinkage: est. R* (Bayes α)* row, shrinkage effectively eliminates the loss relative to the recursive even if the estimate of the rolling window isn't conditioned on the statistical significance of the break test statistic.

Using Bayesian model averaging to combine recursive and rolling forecasts can also essentially match the recursive forecast in average accuracy, if a large prior weight is placed on the recursive model. With the large prior on the recursive forecast, on average the MSE of the BMA forecast exceeds the MSE of the recursive projection by no more than .2 or .3 percent. But with all models having equal weight in the prior, the BMA forecast is somewhat less accurate, exceeding the recursive MSE by between 2 and 3 percent, depending on the DGP and forecast sample. For example, with DGP 2-S and a forecast sample of 40 observations ($\lambda_P = .4$), the *BMA, equal prior prob.* forecast has an average MSE 3.1 percent larger than the recursive forecast.

4.3.2 DGPs with Breaks: Average MSEs

For the breaks imposed in our DGPs, the theoretical results in section 2 imply that, in population and within the class of forecasts without any combination or shrinkage, predictions based on a rolling window of the known R_t^* (α=0) observations will have the lowest MSE. The Monte Carlo results in Tables 3 and 4 show that the result carries over to finite samples. Moreover, in most but not all cases, the known R_t^* (α=0) forecast has MSE lower than the Bayesian shrinkage forecast based on the known R_t^* (Bayes α). For example, Table 3 reports that, for DGP 1-B1, $\lambda_P = .2$, and $\lambda_B = .8$ (a break at observation 80), the known R_t^* (α=0) forecast has an average MSE ratio of .874, compared to MSE ratios of .956 for the *rolling: sup Wald R* forecast and .947 for the shrinkage forecast with the known R_t^* (Bayes α) observations. But in some unreported experiments with smaller or longer-ago breaks, the Bayes shrinkage forecast based on the known R_t^* (Bayes α) slightly beats the known R_t^* (α=0) forecast. The ranking of the two approaches can change because, as the break gets smaller, the R_t^* (α=0) window tends to become the recursive window, while R_t^* (Bayes α) remains the window of just post-break observations.

Within the class of feasible approaches, if the timing is just right, a rolling window of arbitrary, fixed size can produce the lowest average MSE. But if the timing is not just right, a simple rolling approach can be inferior to recursive estimation. Consider, for example, Table 3's results for DGP 1-B1. With the break occurring at observation 80 (λ_B = .8), and forecasts constructed for 40 periods (for observations 101 through 140; λ_P = .4), using a rolling window of 20 observations yields an average MSE ratio of .945. But with the break occurring further back in history, at observation 60 (λ_B = .6), rolling estimation with 20 observations yields an average MSE that is 1.1 percent larger than the recursive forecast's. In general, of course, the gain from using a rolling window shrinks as the break moves further back in history.

Overall, the results in Tables 3 and 4 indicate that estimation with an arbitrary rolling window of 40 observations performs pretty well in DGPs with breaks. When the recursive forecast can be beaten, this simple rolling approach often does so, but when little gain can be had from any of the methods considered, rolling forecasts based on 40 observations are not much worse than the recursive. The upper panel of Figure 1, which breaks the forecast period into short, overlapping segments, highlights the potential for an arbitrary window of 40 observations to yield forecasts as accurate as those from the known R_t^* (α=0) approach when the break occurred relatively recently, but to converge toward the performance of the recursive forecast as the break becomes more distant. In particular, Figure 1 presents the average, across Monte Carlo draws, of MSEs computed for rolling five-observation forecast windows (forecast observations 101 through 105, 102 through 106, etc.).

The performance of forecasts with rolling windows determined by formal break tests is somewhat mixed, reflecting the mixed success of the break tests in correctly identifying breaks. For DGPs with relatively large, recent breaks, the reverse CUSUM and sup Wald-based rolling forecasts are slightly to modestly more accurate than recursive forecasts. For example, Table 3 shows that with DGP 1-B1, $\lambda_B = .8$, and $\lambda_P = .4$, the MSE ratios for these two forecasts are .958 and .941, respectively. But, as might be expected, gains tend to shrink or become losses as the break becomes smaller. For DGP 1-B2, the same forecast approaches have MSE ratios of .973 and .991 when $\lambda_B = .8$ and $\lambda_P = .4$ (Table 4). In broad terms, forecasts based on the estimated rolling window R_t^* (α=0) seem to usually perform as well as or slightly better than the reverse CUSUM and sup Wald forecasts. For instance, with $\lambda_B = .8$, and $\lambda_P = .4$, the estimated R_t^* (α=0) forecast has an average MSE ratio of .931 for DGP 1-B1 (Table 3) and .978 for DGP 1-B2 (Table 4).

Nonetheless, the results in Tables 3 and 4 consistently indicate there is some benefit to simple Bayesian shrinkage of estimates based on rolling data samples. In general, apart from those cases in which an arbitrary rolling window is timed just right so as to yield the best feasible forecast, Bayesian shrinkage seems to improve rolling-window forecasts. In terms of average MSE, the shrinkage forecasts are always as good as or better than the recursive forecast. Moreover, some form of a shrinkage-based forecast usually comes close to yielding the maximum gain possible, among the approaches considered. For example, one of the simplest possible approaches, shrinking rolling estimates based on a window of 40 observations, yields MSE ratios of roughly .96 for both DGP 1-B1 and DGP 2-B1 when $\lambda_B = .8$ or .6 (Table 3).

Bayesian shrinkage of the sup Wald-determined rolling estimates (the *shrinkage: sup Wald R* approach) also yields MSE ratios of roughly .96 in these cases. Perhaps even better is the approach of applying Bayesian shrinkage to a rolling estimate based on a sample window of size determined without conditioning on the significance of the break test (the *shrinkage: est. R* (Bayes α)* approach). In the same cases, this approach yields an MSE ratio of about .945.

To the extent that a simple rolling approach can yield a more accurate forecast than the shrinkage approaches, the Monte Carlo results show that the advantage of the arbitrary approaches diminishes as the forecast period grows longer. The average MSEs for rolling windows of forecast observations presented in the lower panel of Figure 1 confirm that, as the break fades into history, the shrinkage approaches catch up to the simple rolling approach.

Finally, the Monte Carlo results indicate that Bayesian model averaging also yields a consistent benefit that is generally at least as large as that provided by any of the other shrinkage approaches. BMA with an equal prior weight on the recursive and rolling models typically yields a gain in MSE as large as that associated with the known optimal rolling window. In DGP 2-B1, for example, the MSE ratios for the known R_t^* (α=0) and BMA equal prior probability forecasts are .845 and .856, respectively. Not surprisingly, with breaks in the DGP, putting a much larger prior probability on the recursive forecast reduces the benefits of BMA (the advantage of the larger prior being that it sharply reduces the costs of BMA when the DGP is stable): in the same example, the MSE ratio for the BMA large prior probability forecast is .914. But even the large prior probability implementation of BMA seems to perform about as well or better than any other feasible approach to forecasting.

4.3.2 MSE distributions

The limited set of Monte Carlo-based probabilities reported in Table 5 show that the qualitative findings based on average MSEs reflect general differences in the distributions of each forecast's MSE. In the interest of brevity, we report a limited set of probabilities; qualitatively, results are similar for other experiments and settings.

For stable DGPs, in line with the earlier finding that forecasts based on arbitrary rolling windows are on average less accurate than recursive forecasts, the probability estimates in the upper panel of the table indicate that the rolling forecasts are almost always less accurate than recursive forecasts. For example, with DGP 1-S and a forecast sample of 20 observations (λ_P = .2), the probability of a forecast based on a rolling estimation window of 40 observations beating a recursive forecast is only 27.1 percent. Another finding in line with the average MSE results is that shrinkage of rolling estimates significantly reduces the probability of the forecast being less accurate than the recursive. Continuing with the same example, the probability of a shrinkage forecast using a rolling window of 40 observations beating a recursive forecast is 40.2 percent. The table also shows that, in stable DGPs, the break estimate-dependent forecasts tend to perform similarly to the recursive because, with breaks not often found, the break-dependent forecast is usually the same as the recursive forecast (note that the *shrinkage: est . R* (Bayes α)* forecast is an exception because it does not condition on the significance of the break test).

For DGPs with breaks, the probabilities in the lower panel of Table 5 show that while beating the recursive forecast on average usually translates into having a better than 50 percent probability of equaling or beating the recursive forecast, in some cases probability rankings can differ from average MSE rankings. That is, one forecast that produces a smaller average gain (against the recursive) than another sometimes has a higher probability of producing a gain. Perhaps not surprisingly, the reversal of rankings tends to occur with rolling vs. shrinkage forecasts, as shrinkage greatly tightens the MSE distribution. For example, with DGP 1-B1, λ_B = .8, and λ_P = .4, the rolling-40 and shrinkage-40 forecasts have average MSE ratios of .889 and .953, respectively (Table 3). Yet, as reported in the lower panel of Table 5, the probabilities of the rolling-40 and shrinkage-40 forecasts having lower MSE than the recursive are 83.6 and 95.7 percent, respectively.

4.3.3 Summary of simulation results

Not surprisingly, there is a simple tradeoff: methods that forecast most accurately when the DGP has a break tend to fare poorly relative to the recursive approach when the DGP is stable. Accordingly, as long as a forecaster puts some probability on the model of interest being stable, the results seem to favor some form of shrinkage approach. Assigning some probability to the potential for stability implies being cautious in the sense of wanting to not fail to beat a recursive forecast. From this perspective in particular, shrinking estimates based on a rolling window of data seems to be effective and valuable, as does Bayesian model averaging with a large prior on the recursive model. On average, both approaches produce a forecast MSE essentially the same as the recursive MSE when the recursive MSE is best. When there are model instabilities, the shrinkage approaches produce a forecast MSE that often captures most of the gain that can achieved with the methods considered in this paper, and beats the recursive with a high probability. Within the class of shrinkage approaches, using a rolling window of an arbitrary 40 data points seems to work well, as does using a rolling window of length determined by Andrews' (1993) sup Wald test (using either the post-break sample or an optimal R* (Bayes α) observations without conditioning on the break). BMA with a large prior probability on the recursive model seems to perform at least as well as these methods.

5. Application Results

Our evaluation of the empirical performance of the various forecast methods described above follows the spirit of Stock and Watson (1996, 2003a), who document systemic instability in simple time series models for macroeconomic variables. We consider a wide range of applications and a long forecast period (1971 to mid-2003) divided into still-substantial subperiods (1971-85 and 1986-2003). In most cases, other studies have found some evidence of instability in each of the applications we consider. In line with common empirical practice, our presented results are simple RMSEs for one-step ahead forecasts.

5.1 Applications and forecast approach

More specifically, we consider the six applications described below. Appendix 2 provides details on the data samples and sources and forecasting model specifications.

(1) Predicting quarterly GDP growth with lagged growth, an interest rate term spread, and the change in the short-term interest rate. The predictive content of term spreads for output growth has been considered in many studies, ranging from Estrella and Hardouvelis (1991) to Hamilton and Kim (2002). Among others, Stock and Watson (2003a), Estrella, Rodrigues, and Schich (2003), and Clark and McCracken (2003a), have found evidence of instability in the relationship. Kozicki (1997) and Ang, Piazzesi, and Wei (2003) suggest short-term interest rates should also be included.

(2) Forecasting nominal GDP growth using lags of nominal growth and M2 growth. The predictive content of money for real and nominal output has also been considered in many studies; a recent example is Amato and Swanson (2001). Many believe the relationship of output to money growth is plagued by instability.

(3) Predicting monthly growth in industrial production with lagged growth and the Institute of Supply Management's index of manufacturing activity (formerly known as the Purchasing Manager's Index). The ISM index, released at the beginning of each month, is widely viewed as a useful predictor of the industrial production figure released later in the month.

(4) Predicting quarterly core CPI inflation with lagged inflation and the output gap. As indicated by the literature survey in Clark and McCracken (2003b), various Phillips curve specifications are widely used for forecasting inflation. Recent studies of Phillips curves include Stock and Watson (1999), Atkeson and Ohanian (2001), Fisher, Liu and Zhu (2002), and Orphanides and van Norden (2003).

(5) Forecasting monthly excess returns in the S&P 500 using lagged returns, the dividend-price ratio, the 1-month interest rate, and the spread between Baa and Aaa corporate bond yields. Instabilities in models of stock returns have been documented by such studies as Paye and Timmermann (2002), Pesaran and Timmermann (2002), and Rapach and Wohar (2002). Our particular model is patterned after that of Pesaran and Timmermann.

(6) Predicting the quarterly change in the 3-month T-bill rate with the prior quarter's spread between the 6-month and 3-month bill rates, a relation implied by the expectations theory of the term structure. Mankiw and Miron (1986) documented the poor fit of the model to data from 1959 through 1979, linking the apparent failure of the model to the behavior of monetary policy. But Lange, Sack, and Whitesell (2003) find the model does fit data starting in roughly the late 1980s.

In this empirical analysis, we consider the same forecast methods included in the Monte Carlo analysis, with some minor modifications. Rather than consider a range of arbitrary rolling window sizes, we examine forecasts based on just a 10-year window. We also, by necessity, drop consideration of the rolling forecast based on the known R_t^* and the shrinkage forecast using the known Bayesian R_t^*. Finally, in the break analysis, we impose a minimum break segment length of five years of data – 20 quarterly observations or 60 monthly observations. And, in conducting Andrews (1993) tests, we use heteroskedasticity-robust variances in forming the Wald statistics.

5.2 Results

In a broad sense, the application results line up with the Monte Carlo results of Section 4. For example, the simple approach of using an arbitrary rolling window of observations in model estimation can yield the most accurate forecasts when the timing is right (as in the GDP-interest rate and nominal GDP-M2 results for 1986-2003) but inferior forecasts when the timing is not (as in the same application results for 1971-85).

Such broad similarities aside, perhaps the most striking result evident in Table 6 is the difficulty of beating the recursive approach, even in the finance-type applications we consider (the stock return and 3-month interest rate-spread example).[10] Despite the extant evidence of instability in many of the applications considered, the recursive forecast is sometimes the best. Most strikingly, in the inflation-output gap application, none of the alternative approaches yields a forecast RMSE smaller than the recursive, for any of the reported sample periods. Indeed, in

[10] Any gains in the empirical results will naturally appear smaller than in the Monte Carlo results because the empirical results are reported in terms of RMSEs, while the Monte Carlo tables report MSEs.

several cases, the alternative forecasts have RMSEs at least 20 percent larger than the recursive forecast.

Nonetheless, there are some approaches that, in terms of RMSE, usually forecast as well as or better than the recursive approach. And, in line with the Monte Carlo results, it is the shrinkage-based forecasts that consistently equal or improve on the recursive forecast. In particular, our take on the applications evidence is that, within the class of methods that improve on the recursive when improvement is possible but match the accuracy of the recursive when improvement is not possible, Bayesian shrinkage of 10-year rolling window estimates performs best. Some of the other methods, such as Bayesian model averaging or discounted least squares, can offer larger gains over the recursive in some periods, but perform poorly when the recursive forecast is best. One of the more dramatic examples of the potential to do poorly when the recursive is best is provided by the core inflation-output gap application, in which no method beats the recursive approach; the Bayesian shrinkage of 10 year estimates simply minimizes the loss, by essentially matching the recursive performance, with RMSE ratios of 1.009 in all sample periods.

Consider, for example, the 3-month interest rate-term spread application. For 1971-85 the 10-year shrinkage forecast is essentially as accurate as the top-ranked recursive forecast, with a RMSE ratio of 1.002. For 1986-2003, the 10-year shrinkage forecast's RMSE ratio is .978, compared to the best RMSE ratio of .948 provided by a simple rolling forecast. A shrinkage forecast that uses a rolling sample window estimated according to (9) doesn't perform as well: the *shrinkage: est. R*(Bayes α)* RMSE ratios are 1.005 and 1.039 for 1971-85 and 1986-2003, respectively. Bayesian model averaging also doesn't perform as well, yielding RMSE ratios of 1.004 and .991 when a large prior weight is placed on the recursive model. In this application, discounted least squares performs as well as shrinkage of rolling estimates based on 10 years of data, with RMSE ratios of 1.006 and .972 for 1971-85 and 1986-2003, respectively.

Similarly, in the GDP-interest rates application, the 10-year shrinkage forecast essentially matches the accuracy of the recursive forecast for 1971-85, with a RMSE ratio of .996; for 1986-2003, the shrinkage forecast's RMSE ratio is .946, compared to the simple 10 year rolling forecast's RMSE of .944. The shrinkage forecast that uses a rolling sample window estimated according to (9) (the *shrinkage: est. R*(Bayes α)* forecast) yields RMSE ratios of .993 and .991 for 1971-85 and 1986-2003, respectively. In this application, Bayesian model averaging performs roughly as well as 10-year shrinkage. For instance, for 1986-2003, BMA with equal prior weight on all models yields an RMSE ratio of .949; BMA with a large prior weight on the recursive yields an RMSE ratio of .977. Discounted least squares also performs well, with RMSE ratios of 1.000 for 1971-85 and .922 for 1986-2003. As this application clearly shows, in some instances Bayesian model averaging and discounted least squares can perform as well as simple shrinkage of 10-year rolling estimates. The advantage of the simple shrinkage approach seems to come in other applications, such as the inflation and stock return cases, in those samples in which no method really beats the recursive approach.

Still other approaches generally don't seem to fare as well as shrinkage. For example, the reverse CUSUM-based forecast, which is almost always based on very short samples, is typically always less accurate than the recursive forecast, sometimes by large margins (for example, in 1986-2003 GDP forecasts). Finally, like some other forecasts, predictions based on a rolling window of an estimated R_t^* (α=0) observations are sometimes more accurate than recursive forecasts (as in the GDP-interest rates and nominal GDP-M2 forecasts for 1986-2003), but sometimes significantly less accurate (as with stock returns).

6. Conclusion

Within this paper we provide several new results that can be used to improve forecast accuracy in an environment characterized by heterogeneity induced by structural change. These methods focus on the selection of the observation window used to estimate model parameters and the possible combination of forecasts constructed using the recursive and rolling schemes. We first provide a characterization of the bias-variance tradeoff that a forecasting agent faces when deciding which of these methods to use. Given this characterization we establish pointwise optimality results for the selection of both the observation window and any combining weights that might be used to construct forecasts.

Overall, the results in the paper suggest a clear benefit – in theory and practice – to some form of combination of recursive and rolling forecasts. Our Monte Carlo results and results for wide range of applications show that shrinking coefficient estimates based on a rolling window of data seems to be effective and valuable. On average, the shrinkage produces a forecast MSE essentially the same as the recursive MSE when the recursive MSE is best. When there are model instabilities, the shrinkage produces a forecast MSE that often captures most of the gain that can achieved with the methods considered in this paper, and beats the recursive with a high probability. Thus, in practice, combining recursive and rolling forecasts – and doing so easily, in the case of Bayesian shrinkage – yields forecasts that are highly likely to be as good as or better than either recursive forecasts or pure rolling forecasts based on an arbitrary, fixed window size.

Appendix 1: General Theoretical Results on the Bias-Variance Tradeoff

In this appendix we provide a theorem that is used to derive Corollaries 2.1, 2.2, 3.1 and 3.2 in the text. A proof of the Theorem is provided in a not-for-publication technical appendix, Clark and McCracken (2004). In the following let $U_{T,t} = (h'_{T,t+\tau}, vec(x_{T,t}x'_{T,t})')'$, $V = \sum_{j=-\tau+1}^{\tau-1}\Omega_{11,j}$ where $\Omega_{11,j}$ is the upper block-diagonal element of Ω_j defined below, $\Rightarrow$ denotes weak convergence, $B^{-1} = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E(x_{T,t}x'_{T,t})$, and W(.) denotes a standard $(k\times 1)$ Brownian motion.

Assumption 1: (a) The DGP satisfies $y_{T,t+\tau} = x'_{T,t}\beta^*_{T,t} + u_{T,t+\tau} = x'_{T,t}\beta^* + T^{-1/2}x'_{T,t}g(t/T) + u_{T,t+\tau}$ for all t, (b) For $s \in (0, 1+\lambda_P]$ $g(t/T) \Rightarrow g(s)$ a nonstochastic square integrable function.

Assumption 2: The parameters are estimated using OLS.

Assumption 3: (a) $T^{-1}\sum_{t=1}^{[rT]} U_{T,t}U'_{T,t-j} \Rightarrow r\Omega_j$ where $\Omega_j = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E(U_{T,t}U'_{T,t-j})$ all $j \geq 0$, (b) $\Omega_{11,j} = 0$ all $j \geq \tau$, (c) $\sup_{T\geq 1, t\leq T+P} E|U_{T,t}|^{2q} < \infty$ some $q > 1$, (d) The zero mean triangular array $U_{T,t} - EU_{T,t} = (h'_{T,t+\tau}, vec(x_{T,t}x'_{T,t} - Ex_{T,t}x'_{T,t})')'$ satisfies Theorem 3.2 of De Jong and Davidson (2000).

Assumption 4: For $s \in (1, 1+\lambda_P]$, (a) $R_t/T \Rightarrow \lambda_R(s) \in (0,s]$, (b) $\alpha_t \Rightarrow \alpha(s) \in [0,1]$, (c) $P/T \rightarrow \lambda_P \in (0,\infty)$.

Theorem 1: Given Assumptions 1 – 4, $\sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau} - \hat{u}^2_{W,t+\tau}) \rightarrow_d$

$$\begin{aligned}
&\{-2\int_1^{1+\lambda_P}(1-\alpha(s))[s^{-1}W(s)-\lambda_R^{-1}(s)(W(s)-W(s-\lambda_R(s)))]'V^{1/2}BV^{1/2}dW(s)\\
&\quad + \int_1^{1+\lambda_P}(1-\alpha^2(s))s^{-2}W(s)'V^{1/2}BV^{1/2}W(s)ds\\
&\quad - \int_1^{1+\lambda_P}(1-\alpha(s))^2\lambda_R^{-2}(W(s)-W(s-\lambda_R(s)))'V^{1/2}BV^{1/2}(W(s)-W(s-\lambda_R(s)))]ds\ \}\\
&\quad - 2\int_1^{1+\lambda_P}\alpha(s)(1-\alpha(s))s^{-1}\lambda_R^{-1}(s)W(s)'V^{1/2}BV^{1/2}(W(s)-W(s-\lambda_R(s)))ds\ \}\\
&+ 2\{-\int_1^{1+\lambda_P}(1-\alpha(s))[s^{-1}(\int_0^s g(r)dr)-\lambda_R^{-1}(s)(\int_{s-\lambda_R(s)}^s g(r)dr)]'V^{1/2}dW(s)\\
&\quad + \int_1^{1+\lambda_P}[(1-\alpha^2(s))s^{-2}W(s)'V^{1/2}(\int_0^s g(r)dr)-(1-\alpha)^2\lambda_R^{-2}(s)(W(s)-W(s-\lambda_R(s)))'V^{1/2}(\int_{s-\lambda_R(s)}^s g(r)dr)]ds\\
&\quad - \int_1^{1+\lambda_P}\alpha(s)(1-\alpha(s))s^{-1}\lambda_R^{-1}(s)W(s)'V^{1/2}(\int_{s-\lambda_R(s)}^s g(r)dr)ds\\
&\quad - \int_1^{1+\lambda_P}\alpha(s)(1-\alpha(s))s^{-1}\lambda_R^{-1}(s)(W(s)-W(s-\lambda_R(s)))'V^{1/2}(\int_0^s g(r)dr)]ds\\
&\quad - \int_1^{1+\lambda_P}(1-\alpha(s))g(s)'V^{1/2}[s^{-1}W(s)-\lambda_R^{-1}(s)(W(s)-W(s-\lambda_R(s)))]ds\ \}\\
&+ \{-2\int_1^{1+\lambda_P}(1-\alpha(s))g(s)'B^{-1}[s^{-1}(\int_0^s g(r)dr)-\lambda_R^{-1}(s)(\int_{s-\lambda_R(s)}^s g(r)dr)]ds\\
&\quad + \int_1^{1+\lambda_P}[(1-\alpha^2(s))s^{-2}(\int_0^s g(r)dr)'B^{-1}(\int_0^s g(r)dr)-(1-\alpha(s))^2\lambda_R^{-2}(\int_{s-\lambda_R(s)}^s g(r)dr)'B^{-1}(\int_{s-\lambda_R(s)}^s g(r)dr)]ds\\
&\quad - 2\int_1^{1+\lambda_P}\alpha(s)(1-\alpha(s))s^{-1}\lambda_R^{-1}(s)(\int_0^s g(r)dr)'B^{-1}(\int_{s-\lambda_R(s)}^s g(r)dr)ds\ \}
\end{aligned}$$

$$= \int_1^{1+\lambda_P}\xi_W(s) = \{\int_1^{1+\lambda_P}\xi_{W1}(s)\} + \{\int_1^{1+\lambda_P}\xi_{W2}(s)\} + \{\int_1^{1+\lambda_P}\xi_{W3}(s)\}.$$

Appendix 2: Application Details

Unless otherwise noted, all data are taken from the FAME database of the Board of Governors and end in 2003:Q2 (quarterly data) or June 2003 (monthly data). All growth rates and inflation rates are calculated as log changes. Note that, while omitted in the table listing, in all cases the forecasting model includes a constant in the set of predictors. In the table, *start point* refers to the beginning of the regression sample, determined by the availability of the raw data, any differencing, and lag orders.

application	predictand (data frequency)	predictors	data notes
1. GDP-interest rates	real GDP growth (qly)	one lag of: GDP growth; 10 year Treasury bond yield less the 3 month T-bill rate; and the change in the T-bill rate.	start point: 1953:Q3.
2. Nominal GDP-M2	nominal GDP growth (qly)	two lags of: nominal GDP growth and M2 growth	start point: 1959:Q4
3. IP-ISM	growth in industrial production (mly)	one lag of IP growth and the current value of the index of the Institute of Supply Management	start point: January 1948
4. Inflation-output gap	core CPI inflation (qly)	four lags of core inflation and one lag of the output gap, defined as log(GDP/CBO's potential GDP)	start point: 1958:Q2 CBO's potential output taken from the St. Louis Fed's FRED database
5. Stock returns	excess return, S&P 500 (mly)	one lag of: excess return; dividend-price ratio; 1-month nominal interest rate less average over past 12 months; and Baa – Aaa yield spread	start point: February 1954 end point: June 2002 excess return = return less 1-month interest rate, where *return* = $(p_t + d_t)/p_{t-1} - 1$,[11] and the price is from the last business day dividend-price ratio = (average of dividends from $t\text{-}11$ to t)/ p_t data sources: (1) S&P 500 dividend data taken from Robert Shiller's web page. (2) 1-month interest rate taken from Compustat database.
6. 3-mo. Interest rate-term spread	change in 3-month T-bill rate (qly)	one lag of the spread between the 6-month and 3-month T-bill rates	start point: 1959:Q2 quarterly values are the interest rates on the last day of the quarter

[11] Note that the d_t that enters the return calculation is Shiller's reported dividend series divided by 12.

References

Amato, J.D. and N.R. Swanson (2001): "The Real-Time Predictive Content of Money for Output," *Journal of Monetary Economics*, 48, 3-24.

Andrews, D.W.K. (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-56.

Ang, A., M. Piazzesi and M. Wei (2003): "What Does the Yield Curve Tell Us About GDP Growth?," manuscript, Columbia University.

Atkeson, Andrew, and Lee E. Ohanian (2001): "Are Phillips Curves Useful for Forecasting Inflation?," *Quarterly Review*, Federal Reserve Bank of Minneapolis, 25, 2-11.

Bai, J. (1997): "Estimation of a Change Point in Multiple Regression Models," *Review of Economics and Statistics*, 79, 551-63.

Bai, J. and P. Perron (2003): "Computation and Analysis of Multiple Structural-Change Models," *Journal of Applied Econometrics*, 18, 1-22.

Clark, T.E. and M.W. McCracken (2003a): "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics*, forthcoming.

Clark, T.E. and M.W. McCracken (2003b): "The Predictive Content of the Output Gap for Inflation: Resolving In-Sample and Out-of-Sample Evidence," Research Working Paper 03-06, Federal Reserve Bank of Kansas City.

Clark, T.E. and M.W. McCracken (2004): "Technical Appendix to 'Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts'," manuscript, University of Missouri-Columbia.

Durbin, J. (1969): "Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least Squares Residuals," *Biometrika* 56, 1-15.

Edgerton, D. and C. Wells (1994): "Critical Values for the CUSUMSQ Statistic in Medium and Large Sized Sample," *Oxford Bulletin of Economics and Statistics*, 56, 355-65.

Elliott, G. and A. Timmermmann (2003): "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions," *Journal of Econometrics*, forthcoming.

Estrella, A. and G.A. Hardouvelis (1991): "The Term Structure as a Predictor of Real Economic Activity," *Journal of Finance*, 46, 555-76.

Estrella, A., A.P. Rodrigues and S. Schich (2003): "How Stable is the Predictive Power of the Yield Curve? Evidence from Germany and the United States," *Review of Economics and Statistics*, 85, 629-44.

Fisher, J.D.M, C.T. Liu, and R. Zhou (2002): "When Can We Forecast Inflation?" *Economic Perspectives*, Federal Reserve Bank of Chicago (First Quarter), 30-42.

Giacomini, R. and H. White (2003): "Tests of Conditional Predictive Ability", manuscript, UCSD.

Hamilton, J.D. and D.H. Kim (2002): "A Re-Examination of the Predictability of Economic Activity Using the Yield Spread," *Journal of Money, Credit, and Banking*, 34, 340-60.

Hansen, B.E. (1992): "Convergence to Stochastic Integrals for Dependent Heterogeneous Processes, *Econometric Theory* 8 (December), 489-500.

Hansen, B.E. (1997): "Approximate Asymptotic *P* Values for Structural-Change Models," *Journal of Business and Economic Statistics*, 15, 60-67.

Hansen, B.E. (2000): "Testing for Structural Change in Conditional Models," *Journal of Econometrics* 97 (July), 93-116.

Inoue, A. and B. Rossi (2003): "Recursive Predictability Tests for Real-Time Data," manuscript, Duke University.

Koop, G. and S. Potter (2003): "Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging," manuscript, Federal Reserve Bank of New York.

Kozicki, S. (1997): "Predicting Real Growth and Inflation with the Yield Spread," *Economic Review*, Federal Reserve Bank of Kansas City, Fourth Quarter, 39-57.

Lange, J., B. Sack, and W. Whitesell (2003): "Anticipations of Monetary Policy in Financial Markets," *Journal of Money, Credit, and Banking*, 35 (Dec.), 889-909.

Mankiw, N.G. and J.A. Miron (1986): "The Changing Behavior of the Term Structure of Interest Rates," *Quarterly Journal of Economics*, 101 (May), 211-28.

Min, C. and A. Zellner (1993): "Bayesian and non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates", *Journal of Econometrics*, 56, 89-118.

Orphanides, A. and S. van Norden (2003): "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," Scientific Series 2003s-01, CIRANO, January.

Paye, B.S. and A. Timmermann (2002): "How Stable Are Financial Prediction Models? Evidence from US and International Stock Market Data," manuscript, UCSD.

Pesaran, M.H. and A. Timmermann (2002a): "Market Timing and Return Prediction Under Model Instability," *Journal of Empirical Finance*, 19, 495-510.

Pesaran, M.H. and A. Timmermann (2002b): "Model Instability and Choice of observation window," manuscript, UCSD.

Rapach, D.E., and M.E. Wohar (2002): "Structural Change and the Predictability of Stock Returns," manuscript, St. Louis University.

Stock, J.H. and M.W. Watson (1996): "Evidence on Structural Stability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14, 11-30.

Stock, J.H., and M.W. Watson (1999): "Forecasting Inflation," *Journal of Monetary Economics* 44, 293-335.

Stock, J.H. and M.W. Watson (2003a): "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 41, 788-829.

Stock, J.H. and M.W. Watson (2003b): "Combination Forecasts of Output Growth in a Seven–Country Data Set", manuscript, Harvard University and Princeton University.

West, K.D. (1996): "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067-84.

Wright, J.H., (2003): "Forecasting U.S. Inflation by Bayesian Model Averaging," manuscript, Board of Governors of the Federal Reserve System.

Yao, Y-C. (1988): "Estimating the Number of Change-Points Via Schwarz' Criterion," *Statistics and Probability Letters*, 6, 181-89.

Table 1: Summary of Forecast Approaches

approach	description of coefficient estimates
recursive	use all available data
rolling: R=20	use 20 most recent observations
rolling: R=40	use 40 most recent observations
rolling: R=60	use 60 most recent observations
shrinkage: R=20	estimates based on 20 most recent observations shrunken toward recursive model estimates, using (7)
shrinkage: R=40	estimates based on 40 most recent observations shrunken toward recursive model estimates, using (7)
shrinkage: R=40	estimates based on 60 most recent observations shrunken toward recursive model estimates, using (7)
rolling: reverse CUSUM R	use data since break identified by reverse order CUSUM (1% sig.level)
rolling: sup Wald R	use data since break identified by Andrews (1993) sup Wald test for a single break (5% sig.level)
shrinkage: sup Wald R	sup Wald-based coefficient estimates shrunken toward recursive model estimates, using (7)
rolling: known R* (α=0)	using R* most recent observations, where R* is determined using (4) and the known values of the break point, the break size, and the population moments as specified in the DGP
rolling: estimated R* (α=0)	using R* most recent observations, where R* is estimated using (4) and sup Wald-based estimates of the break point and size and sample moment estimates.
shrinkage: known R* (Bayes α)	shrinkage of rolling coefficient estimates based on Bayesian R* observations – determined using (9) and the known features of the DGP – toward recursive estimates
shrinkage: est. R* (Bayes α)	Bayesian shrinkage of rolling coefficient estimates using estimate of Bayesian R* observations based on (9), using Wald-based estimates of the break point and size and sample moment estimates – but regardless of the break test's significance
BMA, equal prior prob.	Bayesian model averaging of recursive and rolling forecasts, with rolling forecasts using each possible start date between observations 20 and t-20. The prior probability on each model is 1/number of models. The shrinkage coefficient ϕ = .2.
BMA, large prior prob.	Same as above, except that the prior probability on the recursive model is .7 and the prior on each rolling model is .3/number of models.
DLS	Discounted least squares with a discount rate of .99.

Table 2: Monte Carlo Results for Stable DGPs, Average MSEs

(average MSE for recursive, and ratio of average MSE to recursive average for other forecasts)

	DGP 1-S				DGP 2-S			
	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1
recursive	1.029	1.030	1.023	1.022	1.029	1.022	1.020	1.020
rolling: R=20	1.152	1.159	1.165	1.170	1.202	1.207	1.211	1.215
rolling: R=40	1.052	1.056	1.060	1.062	1.066	1.071	1.074	1.078
rolling: R=60	1.024	1.026	1.029	1.032	1.029	1.035	1.036	1.040
shrinkage: R=20	1.001	1.001	1.002	1.002	1.000	1.001	1.001	1.000
shrinkage: R=40	1.003	1.003	1.003	1.002	1.000	1.001	1.001	1.001
shrinkage: R=60	1.002	1.002	1.002	1.002	1.001	1.002	1.002	1.002
rolling: reverse CUSUM R	1.004	1.014	1.023	1.037	1.005	1.017	1.028	1.047
rolling: sup Wald R	1.011	1.014	1.014	1.013	1.033	1.031	1.030	1.027
shrinkage: sup Wald R	1.000	1.000	1.000	1.000	1.001	1.001	1.001	1.001
rolling: known R* (α=0)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
rolling: estimated R* (α=0)	1.008	1.010	1.010	1.009	1.027	1.024	1.023	1.021
shrinkage: known R*(Bayes α)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
shrinkage: est. R*(Bayes α)	1.005	1.005	1.005	1.005	1.003	1.004	1.003	1.003
BMA, equal prior prob.	1.024	1.024	1.024	1.021	1.031	1.031	1.028	1.025
BMA, large prior prob.	1.002	1.002	1.002	1.002	1.002	1.003	1.003	1.002
DLS	1.008	1.010	1.011	1.013	1.006	1.008	1.009	1.011

Notes:

1. DGPs DGP 1-S and DGP 2-S are defined in Section 4.1.
2. The total number of observations generated for each experiment is 200. Forecasting begins with observation 101. Results are reported for forecasts evaluated from period 101 through $(1+\lambda_P)100$.
3. The forecast approaches listed in the first column are defined in Table 1.
4. The table entries are based on averages of forecast MSEs across 1000 Monte Carlo simulations. For the recursive forecast, the table reports the average MSEs. For the other forecasts, the table reports the ratio of the average MSE to the average recursive MSE.

Table 3: Baseline Monte Carlo Results for DGPs with Breaks, Average MSEs

(average MSE for recursive, and ratio of average MSE to recursive average for other forecasts)

Break point: λ_B = .8

	DGP 1-B1				DGP 2-B1			
	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1
recursive	1.279	1.254	1.221	1.185	1.310	1.284	1.262	1.231
rolling: R=20	0.922	0.945	0.969	1.002	0.915	0.931	0.948	0.975
rolling: R=40	0.893	0.889	0.902	0.924	0.881	0.868	0.875	0.893
rolling: R=60	0.936	0.912	0.909	0.919	0.932	0.902	0.888	0.890
shrinkage: R=20	0.961	0.966	0.971	0.977	0.966	0.971	0.975	0.981
shrinkage: R=40	0.957	0.953	0.957	0.964	0.959	0.956	0.959	0.966
shrinkage: R=60	0.973	0.962	0.958	0.961	0.973	0.964	0.960	0.962
rolling: reverse CUSUM R	0.991	0.958	0.946	0.952	0.992	0.954	0.931	0.929
rolling: sup Wald R	0.956	0.941	0.937	0.937	0.932	0.909	0.899	0.893
shrinkage: sup Wald R	0.965	0.958	0.956	0.955	0.963	0.956	0.953	0.951
rolling: known R* (α=0)	0.874	0.874	0.881	0.895	0.853	0.845	0.847	0.857
rolling: estimated R* (α=0)	0.944	0.931	0.929	0.930	0.926	0.904	0.893	0.890
shrinkage: known R*(Bayes α)	0.947	0.944	0.944	0.947	0.951	0.947	0.947	0.948
shrinkage: est. R*(Bayes α)	0.947	0.943	0.944	0.947	0.951	0.947	0.946	0.947
BMA, equal prior prob.	0.880	0.878	0.885	0.898	0.865	0.856	0.857	0.864
BMA, large prior prob.	0.933	0.926	0.927	0.931	0.925	0.914	0.909	0.907
DLS	0.928	0.918	0.917	0.921	0.934	0.925	0.920	0.915

Break point: λ_B = .6

	DGP 1-B1				DGP 2-B1			
	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1
recursive	1.188	1.173	1.148	1.125	1.217	1.199	1.186	1.165
rolling: R=20	0.993	1.011	1.030	1.055	0.986	0.997	1.010	1.031
rolling: R=40	0.910	0.925	0.941	0.963	0.888	0.900	0.912	0.932
rolling: R=60	0.911	0.912	0.924	0.942	0.896	0.890	0.895	0.909
shrinkage: R=20	0.974	0.977	0.981	0.985	0.979	0.982	0.985	0.988
shrinkage: R=40	0.954	0.960	0.965	0.973	0.957	0.963	0.967	0.974
shrinkage: R=60	0.958	0.955	0.959	0.967	0.958	0.957	0.960	0.966
rolling: reverse CUSUM R	0.982	0.958	0.960	0.974	0.979	0.944	0.938	0.950
rolling: sup Wald R	0.947	0.944	0.947	0.952	0.914	0.908	0.906	0.910
shrinkage: sup Wald R	0.958	0.957	0.958	0.962	0.953	0.952	0.952	0.954
rolling: known R* (α=0)	0.896	0.904	0.913	0.925	0.870	0.872	0.877	0.888
rolling: estimated R* (α=0)	0.937	0.936	0.941	0.947	0.906	0.902	0.901	0.906
shrinkage: known R*(Bayes α)	0.946	0.948	0.951	0.956	0.948	0.948	0.949	0.952
shrinkage: est. R*(Bayes α)	0.946	0.947	0.951	0.956	0.947	0.947	0.948	0.951
BMA, equal prior prob.	0.896	0.904	0.914	0.928	0.873	0.876	0.881	0.892
BMA, large prior prob.	0.930	0.932	0.938	0.946	0.911	0.912	0.913	0.918
DLS	0.928	0.927	0.932	0.941	0.933	0.929	0.928	0.927

Notes:

1. DGPs DGP 1-B1 and DGP 2-B1 are defined in Section 4.1.
2. The total number of observations in each experiment is 200. Forecasting begins with observation 101. Results are reported for forecasts evaluated from period 101 through $(1+\lambda_P)100$. The break in the DGP occurs at observation $\lambda_B 100$.
3. The forecast approaches listed in the first column are defined in Table 1.
4. The table entries are based on averages of forecast MSEs across 1000 Monte Carlo simulations. For the recursive forecast, the table reports the average MSEs. For the other forecasts, the table reports the ratio of the average MSE to the average recursive MSE.

Table 4: Auxiliary Monte Carlo Results for DGPs with Breaks, Average MSEs

(average MSE for recursive, and ratio of average MSE to recursive average for other forecasts)

Break point: λ_B = .8

	DGP 1-B2				DGP 2-B2			
	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1
recursive	1.202	1.181	1.152	1.124	1.196	1.173	1.153	1.127
rolling: R=20	0.988	1.012	1.034	1.064	1.033	1.047	1.065	1.094
rolling: R=40	0.931	0.936	0.951	0.972	0.944	0.945	0.960	0.984
rolling: R=60	0.953	0.940	0.945	0.958	0.955	0.944	0.948	0.961
shrinkage: R=20	0.964	0.970	0.975	0.981	0.962	0.968	0.972	0.979
shrinkage: R=40	0.962	0.961	0.965	0.972	0.959	0.958	0.962	0.969
shrinkage: R=60	0.976	0.969	0.968	0.972	0.974	0.967	0.965	0.969
rolling: reverse CUSUM R	0.991	0.973	0.972	0.985	0.994	0.977	0.977	0.995
rolling: sup Wald R	1.000	0.991	0.990	0.988	1.021	1.006	1.002	0.997
shrinkage: sup Wald R	0.975	0.972	0.972	0.973	0.973	0.969	0.968	0.969
rolling: known R* (α=0)	0.925	0.925	0.932	0.942	0.950	0.939	0.940	0.946
rolling: estimated R* (α=0)	0.986	0.978	0.978	0.978	1.005	0.990	0.987	0.984
shrinkage: known R*(Bayes α)	0.953	0.953	0.956	0.961	0.951	0.951	0.953	0.958
shrinkage: est. R*(Bayes α)	0.955	0.954	0.957	0.962	0.955	0.954	0.955	0.960
BMA, equal prior prob.	0.919	0.921	0.929	0.941	0.925	0.924	0.929	0.940
BMA, large prior prob.	0.953	0.952	0.954	0.960	0.952	0.950	0.951	0.957
DLS	0.939	0.936	0.940	0.949	0.934	0.931	0.933	0.941

Break point: λ_B = 1

	DGP 1-B1				DGP 2-B1			
	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1	λ_P=.20	λ_P=.40	λ_P=.60	λ_P=1
recursive	1.394	1.354	1.308	1.257	1.480	1.411	1.368	1.314
rolling: R=20	0.975	0.942	0.950	0.973	1.009	0.951	0.946	0.958
rolling: R=40	0.978	0.929	0.914	0.917	0.986	0.926	0.901	0.895
rolling: R=60	0.988	0.955	0.931	0.918	0.987	0.954	0.924	0.898
shrinkage: R=20	0.980	0.971	0.972	0.976	0.985	0.976	0.976	0.979
shrinkage: R=40	0.987	0.972	0.965	0.966	0.989	0.975	0.968	0.968
shrinkage: R=60	0.993	0.981	0.971	0.964	0.993	0.982	0.973	0.966
rolling: reverse CUSUM R	0.997	0.975	0.950	0.939	0.998	0.973	0.940	0.921
rolling: sup Wald R	1.000	0.975	0.959	0.941	1.009	0.963	0.934	0.906
shrinkage: sup Wald R	0.995	0.980	0.971	0.962	0.995	0.979	0.969	0.960
rolling: known R* (α=0)	0.984	0.923	0.906	0.899	0.996	0.915	0.888	0.871
rolling: estimated R* (α=0)	0.997	0.969	0.952	0.934	1.005	0.959	0.930	0.903
shrinkage: known R*(Bayes α)	0.980	0.964	0.957	0.952	0.985	0.969	0.961	0.955
shrinkage: est. R*(Bayes α)	0.983	0.965	0.958	0.952	0.987	0.970	0.962	0.955
BMA, equal prior prob.	0.963	0.920	0.906	0.899	0.969	0.916	0.892	0.877
BMA, large prior prob.	0.981	0.957	0.946	0.936	0.969	0.919	0.895	0.879
DLS	0.970	0.944	0.931	0.920	0.976	0.951	0.936	0.921

Notes:

1. DGPs DGP 1-B2, DGP 2-B2, DGP 1-B1, and DGP 2-B1 are defined in Section 4.1.
2. The total number of observations in each experiment is 200. Forecasting begins with observation 101. Results are reported for forecasts evaluated from period 101 through $(1+\lambda_P)100$. The break in the DGP occurs at observation $\lambda_B 100$.
3. The forecast approaches listed in the first column are defined in Table 1.
4. The table entries are based on averages of forecast MSEs across 1000 Monte Carlo simulations. For the recursive forecast, the table reports the average MSEs. For the other forecasts, the table reports the ratio of the average MSE to the average recursive MSE.

Table 5: Monte Carlo Probabilities of Equaling or Beating Recursive MSE

(Stable) DGP 1-S

	λ_P=.20		λ_P=.40		λ_P=.60		λ_P=1	
	Pr(=REC)	*Pr(<REC)*	*Pr(=REC)*	*Pr(<REC)*	*Pr(=REC)*	*Pr(<REC)*	*Pr(=REC)*	*Pr(<REC)*
rolling: R=20	0.000	0.162	0.000	0.071	0.000	0.022	0.000	0.007
rolling: R=40	0.000	0.271	0.000	0.177	0.000	0.109	0.000	0.045
rolling: R=60	0.000	0.343	0.000	0.280	0.000	0.215	0.000	0.125
shrinkage: R=20	0.000	0.417	0.000	0.410	0.000	0.389	0.000	0.384
shrinkage: R=40	0.000	0.402	0.000	0.378	0.000	0.360	0.000	0.360
shrinkage: R=60	0.000	0.425	0.000	0.415	0.000	0.377	0.000	0.358
rolling: reverse CUSUM R	0.000	0.432	0.000	0.335	0.000	0.246	0.000	0.111
rolling: sup Wald R	0.863	0.033	0.795	0.030	0.751	0.020	0.675	0.020
shrinkage: sup Wald R	0.863	0.053	0.795	0.067	0.751	0.079	0.675	0.084
rolling: known R* (α=0)	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
rolling: estimated R* (α=0)	0.863	0.036	0.795	0.036	0.751	0.024	0.675	0.029
shrinkage: known R*(Bayes α)	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
shrinkage: est. R*(Bayes α)	0.000	0.422	0.000	0.393	0.000	0.360	0.000	0.332
BMA, equal prior prob.	0.000	0.311	0.000	0.237	0.000	0.203	0.000	0.151
BMA, large prior prob.	0.000	0.389	0.000	0.347	0.000	0.317	0.000	0.316
DLS	0.000	0.357	0.000	0.306	0.000	0.260	0.000	0.202

(Break) DGP 1-B1, λ_B = .8

	λ_P=.20		λ_P=.40		λ_P=.60		λ_P=1	
	Pr(=REC)	*Pr(<REC)*	*Pr(=REC)*	*Pr(<REC)*	*Pr(=REC)*	*Pr(<REC)*	*Pr(=REC)*	*Pr(<REC)*
rolling: R=20	0.000	0.636	0.000	0.622	0.000	0.572	0.000	0.492
rolling: R=40	0.000	0.773	0.000	0.836	0.000	0.859	0.000	0.845
rolling: R=60	0.000	0.739	0.000	0.862	0.000	0.888	0.000	0.924
shrinkage: R=20	0.000	0.881	0.000	0.934	0.000	0.950	0.000	0.965
shrinkage: R=40	0.000	0.873	0.000	0.957	0.000	0.975	0.000	0.985
shrinkage: R=60	0.000	0.795	0.000	0.934	0.000	0.975	0.000	0.990
rolling: reverse CUSUM R	0.000	0.553	0.000	0.748	0.000	0.785	0.000	0.790
rolling: sup Wald R	0.253	0.448	0.106	0.580	0.063	0.659	0.025	0.745
shrinkage: sup Wald R	0.253	0.619	0.106	0.779	0.063	0.847	0.025	0.920
rolling: known R* (α=0)	0.000	0.764	0.000	0.858	0.000	0.904	0.000	0.941
rolling: estimated R* (α=0)	0.253	0.480	0.106	0.625	0.063	0.704	0.025	0.789
shrinkage: known R*(Bayes α)	0.000	0.923	0.000	0.974	0.000	0.989	0.000	0.995
shrinkage: est. R*(Bayes α)	0.000	0.889	0.000	0.943	0.000	0.969	0.000	0.986
BMA, equal prior prob.	0.000	0.845	0.000	0.930	0.000	0.952	0.000	0.971
BMA, large prior prob.	0.000	0.892	0.000	0.955	0.000	0.975	0.000	0.993
DLS	0.000	0.865	0.000	0.947	0.000	0.971	0.000	0.978

Notes:

1. DGPs DGP 1-S and DGP 1-B1 are defined in Section 4.1.
2. The total number of observations in each experiment is 200. Forecasting begins with observation 101. Results are reported for forecasts evaluated from period 101 through $(1+\lambda_P)100$. The break in the DGP occurs at observation $\lambda_B 100$.
3. The forecast approaches listed in the first column are defined in Table 1.
4. The table entries are frequencies (percentages of 1000 Monte Carlo draws) with which a given forecast approach yields a forecast MSE less than or equal to the recursive forecast's MSE.

Table 6: Application Results

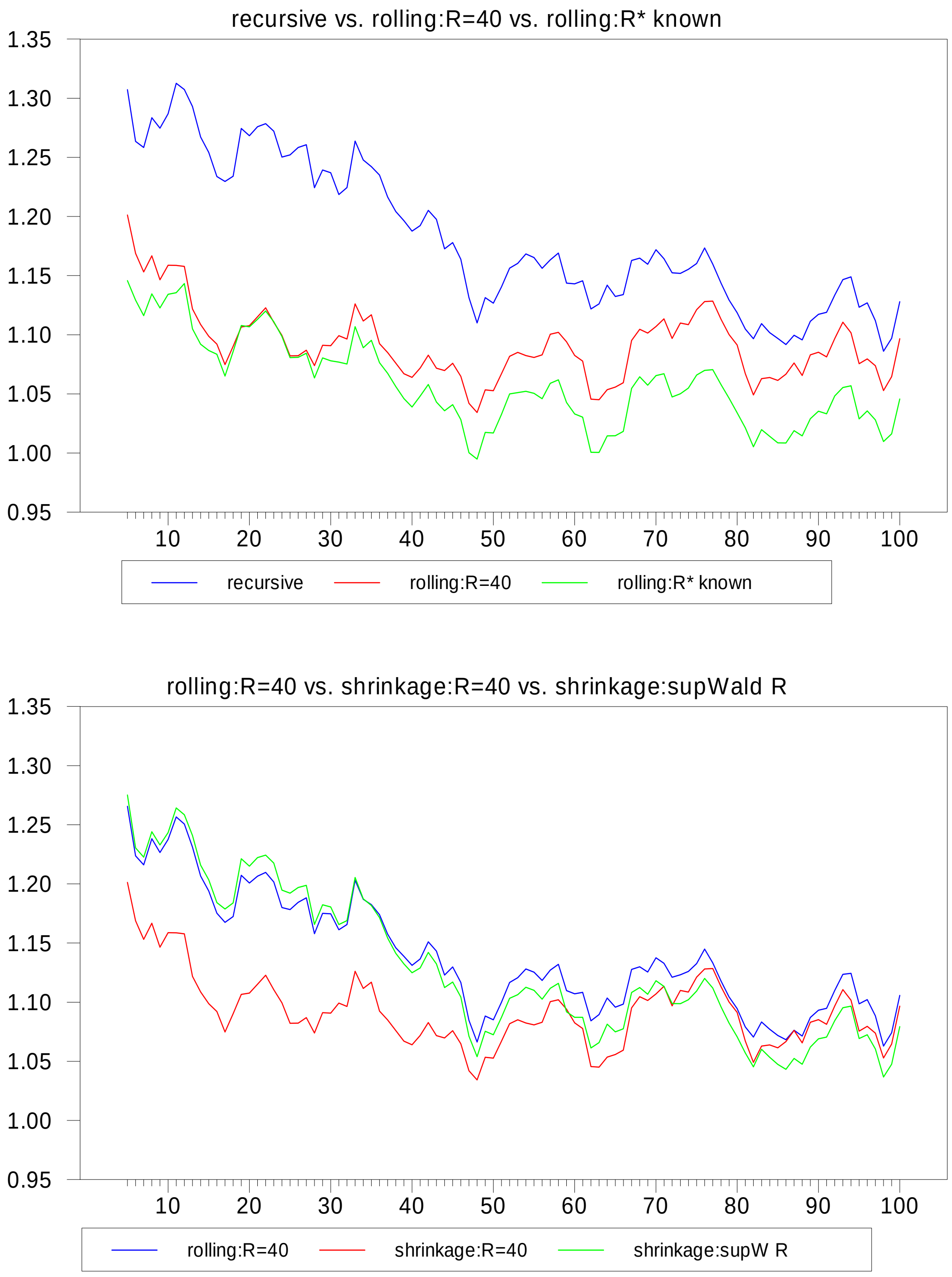
(RMSE for recursive forecast, and ratio of RMSE to recursive RMSE for other forecasts)

	GDP-interest rates			**Nominal GDP-M2**		
	1971-2003	*1971-85*	*1986-2003*	*1971-2003*	*1971-85*	*1986-2003*
recursive	3.323	4.139	2.413	3.597	4.585	2.451
rolling: R=40	0.994	1.013	0.944	0.984	1.023	0.858
shrinkage: R=40	0.982	0.996	0.946	0.986	0.991	0.968
rolling: reverse CUSUM R	1.037	1.016	1.087	1.017	1.022	1.003
rolling: sup Wald R	1.008	1.017	0.987	1.012	1.032	0.947
shrinkage: sup Wald R	0.999	1.002	0.991	0.993	0.999	0.975
rolling: estimated R* (α=0)	1.008	1.017	0.987	1.012	1.032	0.947
shrinkage: est. R*(Bayes α)	0.993	0.993	0.991	0.987	0.991	0.976
BMA, equal prior prob.	0.990	1.005	0.949	0.992	1.014	0.924
BMA, large prior prob.	0.990	0.995	0.977	0.989	0.997	0.967
DLS	0.978	1.000	0.922	0.979	0.999	0.919
	IP-ISM			**Inflation-output gap**		
	1971-2003	*1971-85*	*1986-2003*	*1971-2003*	*1971-85*	*1986-2003*
recursive	8.116	9.977	6.083	1.627	2.310	0.582
rolling: R=40	1.005	1.009	0.995	1.064	1.042	1.327
shrinkage: R=40	0.994	0.994	0.994	1.009	1.009	1.009
rolling: reverse CUSUM R	1.009	1.018	0.986	1.047	1.037	1.173
rolling: sup Wald R	1.006	1.000	1.017	1.249	1.250	1.235
shrinkage: sup Wald R	0.998	0.999	0.993	1.031	1.032	1.020
rolling: estimated R* (α=0)	1.006	1.000	1.017	1.249	1.250	1.235
shrinkage: est. R*(Bayes α)	0.997	0.998	0.995	1.031	1.032	1.020
BMA, equal prior prob.	0.994	0.997	0.988	1.079	1.067	1.233
BMA, large prior prob.	0.996	0.997	0.993	1.024	1.021	1.068
DLS	0.988	0.996	0.969	1.039	1.040	1.023
	Stock returns			**3-mo. Interest rate-term spread**		
	1971-2003	*1971-85*	*1986-2003*	*1971-2003*	*1971-85*	*1986-2003*
recursive	4.523	4.333	4.689	1.186	1.623	0.594
rolling: R=40	1.034	1.037	1.032	1.003	1.011	0.948
shrinkage: R=40	0.999	1.004	0.995	0.999	1.002	0.978
rolling: reverse CUSUM R	1.035	1.048	1.025	0.990	1.004	0.900
rolling: sup Wald R	1.042	1.043	1.042	1.089	1.100	1.013
shrinkage: sup Wald R	0.998	1.005	0.993	1.004	1.005	0.999
rolling: estimated R* (α=0)	1.042	1.043	1.042	1.089	1.100	1.013
shrinkage: est. R*(Bayes α)	1.000	1.007	0.995	1.010	1.005	1.039
BMA, equal prior prob.	1.011	1.030	0.997	1.021	1.026	0.985
BMA, large prior prob.	1.001	1.007	0.997	1.003	1.004	0.991
DLS	1.018	1.049	0.993	1.002	1.006	0.972

Notes:

1. Details of the six applications (data, forecast model specification, etc.) are provided in Appendix 2.
2. The forecast approaches listed in the first column are defined in Table 1.
4. The table entries are based on forecast RMSEs. For the recursive forecast, the table reports the RMSE. For the other forecasts, the table reports the ratio of its RMSE to the recursive RMSE.

Figure 1: Average MSE Across Monte Carlo Draws,
Rolling 5-Period Windows of Forecasts
DGP 1-B1, λ_B = .8



recursive vs. rolling:R=40 vs. rolling:R* known

rolling:R=40 vs. shrinkage:R=40 vs. shrinkage:supWald R