IMPROVING FORECAST ACCURACY BY COMBINING RECURSIVE AND ROLLING FORECASTS

Todd E. Clark and Michael W. McCracken

October 2004

RWP 04-10

Research Division
Federal Reserve Bank of Kansas City

Todd Clark is vice president and economist at the Federal Reserve Bank of Kansas City and Michael McCracken is an assistant professor of Economics at the University of Missouri-Columbia. The authors gratefully acknowledge the excellent research assistance of Taisuke Nakata and helpful comments from Ulrich Muller, Peter Summers, Ken West, Jonathan Wright, seminar participants at the University of Virginia, the Board of Governors and the Federal Reserve Bank of Kansas City, and participants at the following meetings: MEG, Canadian Economic Association, SNDE, MEC, 2004 NBER Summer Institute, NBER/NSF Time Series Conference and the conference for young researchers on Forecasting in Time Series.

Clark email: todd.e.clark@kc.frb.org
McCracken email: mccrackenm@missouri.edu

Abstract

This paper presents analytical, Monte Carlo, and empirical evidence on the effectiveness of combining recursive and rolling forecasts 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.

JEL classification: C53, C12, C52

Keywords: Structural breaks, forecasting, model averaging

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. If the earliest available data follow a data-generating process unrelated to the present then using such data in estimation 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, reducing the sample in order to reduce heterogeneity also increases 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 parameters (e.g. Stock and Watson, 2003). Since both financial and macroeconomic series are known to exhibit 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.

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. Combination could be seen as a form of shrinkage. Min and Zellner (1993), Koop and Potter (2003), Stock and Watson (2003), Wright (2003), Maheu and Gordon

(2004), and Pesaran, Pettenuzzo and Timmermann (2004) have found some form of shrinkage 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 the optimal observation window for the rolling scheme but also a solution for the joint optimal observation window and combining weights.

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. In particular, shrinking coefficient estimates based on a rolling window of data seems to be effective. 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. The first is the very large and resurgent literature on forecast combination, both theoretical (e.g. Elliott and Timmermann, 2004) and empirical (e.g. Stock and Watson, 2003, 2004). 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.[1]

Finally, our work on the optimal choice of observation window builds on Pesaran and Timmermann (2002b). 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, the size of any change in the residual variance, and 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 combination. Section 5 compares the effectiveness of the forecast methods in a range of empirical applications. The final section concludes. Details pertaining to theory and data are presented in Appendixes 1 and 2.

---

[1] In a related approach, Engle and Smith (1999) allow continuous variation in parameters, but make the rate of variation a function of recent errors in the forecasting model. Larger errors provide a stronger signal of a change in parameters.

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 $1-\alpha_t$ 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[2]

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

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

Note that we allow the dependent variable $y_{T,t+\tau}$, the predictors $x_{T,t}$ and the error term $u_{T,t+\tau}$ to depend upon T, the initial forecasting origin. By doing so we allow the time variation in the parameters to influence 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.

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 $(k \times 1)$ vector of potential predictors x_t

---

[2] The parameter $\beta^*_{T,t}$ does not vary with the forecast horizon τ since, in our analysis, τ is treated as fixed.

which may include lagged dependent variables. Forecasts of the scalar $y_{t+\tau}$, $t = T,...T + P$, $\tau \geq 1$, 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 $\hat{\beta}_{R,t} = \arg\min t^{-1}\sum_{s=1}^{t-\tau}(y_{s+\tau} - x_s'\beta)^2$ and $\hat{\beta}_{L,t} =$ $\arg\min R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}(y_{s+\tau} - x_s'\beta)^2$ 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 $\lim_{T\rightarrow\infty} R_t / T = \lambda_R \in (0,1)$. 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 $\lim_{T\rightarrow\infty} P / T = \lambda_P \in (0,\infty)$ 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 $\sum_{t=T}^{T+P}(\hat{u}_{R,t+\tau}^2 - \hat{u}_{L,t+\tau}^2)$, the difference in the (normalized) MSEs of the recursive and rolling forecasts.[3] As detailed in Theorem 1 in Appendix 1, we show that this statistic has an asymptotic distribution that can be decomposed into three terms:

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

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 (via 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

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

focus on the mean (rather than the distribution) of the bias-variance tradeoff when there are either no breaks or a single break.

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

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

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, the recursive scheme is always best.

2.4 The case of a single break

Suppose that a permanent local structural change, of magnitude $T^{-1/2}g(t/T) = T^{-1/2}\Delta\beta$, occurs in the parameter vector β at time $1 \leq T_B \leq t$ where again, $t = T,...T+P$ 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

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

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

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

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

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. The optimal observation window given this tradeoff is provided 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

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

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 arguments of the integrals in parts (a) and (b) of Corollary 2.1 that contribute to the average expected mean square differential over the duration of forecasting. In particular, the results of Corollary 2.2 follow from maximizing

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

with respect to $\lambda_R(s)$ for each s and keeping track of the relevant corner solutions.

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 decreasing in the ratio $\Delta\beta' B^{-1}\Delta\beta / tr(BV)$. 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'. We therefore suggest a transformed formula. 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 ($T^{-1/2}\Delta\beta$) 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.[4]

[4] 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.

$$\hat{R}_t^* = \begin{cases} t & 0 \leq 1 - 2\hat{\delta}_B(1 - \hat{\delta}_B)(\frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})}) \\ \dfrac{2t(1 - \hat{\delta}_B)^2(\frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})})}{2(1 - \hat{\delta}_B)(\frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})}) - 1} & 1 - 2\hat{\delta}_B(1 - \hat{\delta}_B)(\frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})}) < 0 < \dfrac{1}{1 + 2\hat{\delta}_B(1 - \hat{\delta}_B)(\frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})})} \\ t(1 - \hat{\delta}_B) & \dfrac{1}{1 + 2\hat{\delta}_B(1 - \hat{\delta}_B)(\frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})})} \to 0 \end{cases} . \quad (4)$$

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. Doing so allows us to derive closed form solutions for the optimal observation window. Unfortunately, though, local breaks cannot be consistently estimated (Bai (1997)). We therefore simply use global break magnitudes and dates to estimate (inconsistently) the assumed local magnitudes and optimal sample window. However, our Monte Carlo experiments indicate that the primary difficulty is not the inconsistency of our estimate of the optimal observation window; rather, the primary difficulty is break identification and dating. Optimal rolling window (and combination) forecasts that estimate the size of the break using the known date of the break in the DGP perform essentially as well as forecasts using both the known size and date of the break. Not surprisingly, forecast accuracy deteriorates somewhat when both the size and date of the break are estimated. 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 estimates toward recursive estimates. The third is 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 $\hat{u}^2_{W,t+\tau} = (y_{t+\tau} - x_t'\hat{\beta}_{W,t})^2$.

Using Theorem 1 in Appendix 1, we are able to derive not only the optimal observation window for such a forecast, but also the associated optimal combining weight in the presence of a single structural break. 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

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

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

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

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

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

$$(\lambda_R^*(s), \alpha^*(s)) = (s - \lambda_B, \frac{s}{s + (\frac{\Delta\beta' B^{-1} \Delta\beta}{tr(BV)})\lambda_B(s - \lambda_B)}).$$

Corollary 3.2 provides pointwise (jointly) optimal observation windows and combining weights 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 arguments of the integrals in parts (a) and (b) of Corollary 3.1 that contribute to the average expected mean square differential over the duration of forecasting.

In contrast to the optimal observation window result from Corollary 2.2, the joint optimal solution is surprisingly simple. In particular, the optimal strategy is to combine a rolling forecast that uses all post-break observations with a recursive forecast that uses all observations. In other words, the best strategy for minimizing the mean square forecast error in the presence of a structural break is not so much to optimize the observation window, as suggested in Pesaran and Timmermann (2002b), but rather to focus instead on forecast combination.

Comparative statics for the combining weights are straightforward. As the magnitude of the break increases relative to the precision of the parameter estimates, the weight on the recursive scheme decreases. We also obtain the intuitive result that as the time since the break ($(s - \lambda_B)$) increases, we eventually place all weight on the rolling scheme.

Again though, 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 formula in Corollary 3.2.

$$(\hat{R}_t^*, \hat{\alpha}_t^*) = (t(1-\hat{\delta}_B), \frac{1}{1+(\frac{t\Delta\hat{\beta}'\hat{B}^{-1}\Delta\hat{\beta}}{tr(\hat{B}\hat{V})})\hat{\delta}_B(1-\hat{\delta}_B)}). \quad (5)$$

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 mean 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 $B_R(t) = (t^{-1}\sum_{j=1}^{t-\tau} x_j x_j')^{-1}$.[5] 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 mean rolling parameter estimates given this prior:

$$\tilde{\beta}_{W,t} = [tB_R^{-1}(t) + R_t B_L^{-1}(t)]^{-1}[tB_R^{-1}(t)\hat{\beta}_{R,t} + \textstyle\sum_{s=t-\tau-R_t+1}^{t-\tau} x_s y_{s+\tau}]$$

$$= [B_R^{-1}(t) + (R_t/t)B_L^{-1}(t)]^{-1}B_R^{-1}(t)\hat{\beta}_{R,t} + [(t/R_t)B_R^{-1}(t) + B_L^{-1}(t)]^{-1}B_L^{-1}(t)\hat{\beta}_{L,t}. \quad (6)$$

It is clear from the right-hand side of (6) that the parameter estimates are a linear combination of both recursive and rolling parameter estimates. In contrast to the simple combination considered

---

[5] Since we are using data to parameterize the prior, it is perhaps more appropriate to say that we are using an objective (rather than subjective) prior. See Berger and Pericchi (2004) for discussion.

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. In light of the results from Corollary 3.2, we use all post-break observations when constructing the rolling component of the forecast.

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 observation windows. 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). For example, 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, $\mathrm{Prob}(M_i)$, on each model is just 1/the number of models. We also consider the alternative of putting a large prior weight 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, $\mathrm{Prob}(M_i \mid data)$, of each model, we set the prior on the coefficients equal to the recursive estimates.[6] Specifically, at each forecast origin t we calculate the posterior probability of each model M_i using

$$\mathrm{Prob}(M_i|\mathrm{data}) = \frac{\mathrm{Prob}(\mathrm{data}|M_i)\times\mathrm{Prob}(M_i)}{\sum_{\mathrm{i}}\mathrm{Prob}(\mathrm{data}|M_i)\times\mathrm{Prob}(M_i)},$$

where:

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

ϕ = parameter determining the rate of shrinkage toward the prior

p_i = the number of explanatory variables in model i

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

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.

[6] 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.

4. Monte Carlo Results

We use Monte Carlo simulations of 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 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), and 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:

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

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 two different break dates (a single break in each experiment): λ_B = .6 and .8.

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.[7] We set T, the number of observations preceding the first forecast date, to 100, and consider forecast periods of various lengths: λ_P = .2, .4, .6, and 1.0. For each value of λ_P, forecasts are evaluated over the period T 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

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

using variations on the approaches described above. Table 1 details all of the forecast methods. As to the particulars of our analysis, we note the following.

1. Some break testing details: (a) Our tests are based on the full set of forecast model coefficients, in part for simplicity. (b) We impose a minimum segment length of 20 periods.

2. For all but one of the forecasts that rely on break identification, 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. The exception is the *shrinkage: sup Wald R (all)* forecast, which simply uses the estimated break and break date without requiring the break to be statistically significant.

3. Most of our results using break tests are based on the Andrews (1993) test for a single break, with a 5% significance level.[8] We do, however, consider other approaches. One, for

[7] 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.

[8] At each point in time, the asymptotic p-value of the sup Wald test is calculated using Hansen's (1997) approximation. As noted by Inoue and Rossi (2003) in the context of causality testing, repeated tests in such real time analyses with the use of standard critical values will result in spurious break findings. Using adjusted critical values would improve the stable-DGP performance of some of our break test-based methods. But in DGPs with breaks, performance would deteriorate.

which we report results, is the reverse order CUSUM (of squares) method proposed by Pesaran and Timmermann (2002a), which involves searching backward from each forecast date to find the most recent break.[9] 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. Another, not reported in the interest of brevity, is the BIC criterion of Yao (1988) and Bai and Perron (2003). We omit the results for the BIC, which allows for the potential of multiple breaks, because they are comparable to those reported for the single break sup Wald approach. Yet another approach, which we leave for future research, would be Bayesian break identification (e.g.,Wang and Zivot (2000)).

4. Although we have experimented with various values of the BMA parameter ϕ that determines the rate of shrinkage toward the recursive (a smaller value corresponds to more shrinkage) used in calculating the posterior probabilities, we report results for the single value that seems to work best: ϕ = .2.

5. Because many readers seem to find discounted least squares (DLS) to be a natural alternative, and DLS has come to be widely used in macroeconomic models featuring learning (e.g., Cho, Williams, and Sargent (2002)), we include forecasts based on models estimated with a discount rate of .99.

6. Although infeasible in empirical applications, for benchmarking purposes we report results for forecasts based on the optimal weight α_t^* and window R_t^* calculated using the known features of the DGP – the break point, the break size, and the population moments of the data.[10]

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, we only report actual average MSEs for the recursive forecast. For all other forecasts, we report the ratio of a forecast's average MSE to the recursive forecast's average MSE. To capture potential

[9] 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.

[10] 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.

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 weight α_t^* on the recursive forecast is 1 and the rolling window R_t^* (α =0) in (4) is the full sample. Thus, the known optimal combination forecast, the rolling forecast based on the known R_t^* (α =0), and the Bayesian shrinkage forecast based on the known break date 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 (λ_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. For example, with DGP 2-S and a forecast sample of 40 observations (λ_P = .4), the reverse CUSUM forecast has an average MSE 1.7 percent larger than the recursive forecast. For the same DGP and sample, 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. The optimal combination forecast – a weighted average of the recursive and *rolling: sup Wald R* projections, with estimated weights

– performs slightly better than the rolling forecast. Similarly, the rolling forecast based on an estimate of R_t^* (α =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 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 (5%)*) that, at worse, has an average MSE .1 percent larger than the recursive. As indicated by the results in the *shrinkage: sup Wald R (all)* 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.

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 (λ_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, the combined forecast based on the known optimal α_t^* will have the lowest MSE. Within the class of forecasts without any combination, predictions based on a rolling window of the known R_t^* (α=0) observations should have the lowest MSE. The Monte Carlo results in Tables 3 and 4 bear out these analytical implications: the optimal combination forecast always has the lowest average MSE, with the known R_t^* (α=0) forecast second, although sometimes just trivially so. Moreover, in most but not all cases, the known R_t^* (α=0) forecast has a lower MSE than the Bayesian shrinkage forecast based on the known break date. For example, Table 3 reports that, for DGP 1-B1, λ_P = .2, and λ_B = .8 (a break at observation 80), the optimal combination forecast has an average MSE ratio of .854, compared to MSE ratios of .874 for the known R_t^* (α=0) forecast and .947 for the shrinkage forecast with the known break date. But in some unreported experiments with smaller or longer-ago breaks, the Bayes shrinkage forecast based on the known break date 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 the shrinkage forecast is based on the 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 our 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 performance of forecasts relying on 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, λ_B = .8, and λ_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 λ_B = .8 and λ_P = .4 (Table 4). Either combining the recursive and post-break forecasts according to (7) or constructing a forecast with the estimated rolling window R_t^* (α=0) offers some slight improvement over the reverse CUSUM and sup Wald forecasts. For instance, with λ_B = .8 and λ_P = .4, the optimal combination forecast has an average MSE ratio of .928 for DGP 1-B1 (Table 3) and .977 for DGP 1-B2 (Table 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). In their feasible incarnations, the optimal combination and optimal rolling window methods yield virtually the same average MSEs.

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* (*5%*) 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: sup Wald R (all)* approach). In the same cases, this approach yields an MSE ratio of about .945.

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 nearly as large as that associated with the known optimal combination. In DGP 2-B1, for example, the MSE ratios for the known optimal combination and BMA equal prior probability forecasts are .838 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: post-break R (all)* 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. Assuming a desire to be cautious in the sense of wanting to not fail to beat a recursive forecast, shrinking estimates based on a rolling or post-break sample 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.

5. Application Results

To evaluate the empirical performance of the various forecast combination methods, we follow the spirit of Stock and Watson (1996, 2003) in considering a wide range of applications and forecast performance over various periods (1976-89 and 1990-2003). For a number of the applications, other studies have found some evidence of instability. In line with common empirical practice, our presented results are simple RMSEs for one-step ahead forecasts.

5.1 Applications and forecast approach

The predictands in the 12 applications listed below are widely-studied, broad-scope economic indicators for the U.S. and select other industrial economies (see Appendix 2 for details on the data and model specifications).

(1) Predicting quarterly U.S. GDP growth with lagged growth, an interest rate term spread, and the change in the short-term interest rate (examples: Estrella and Hardouvelis (1991), Hamilton and Kim (2002), and Stock and Watson (2003)).

(2) Forecasting quarterly U.S. core CPI inflation with an AR(4) model (Stock and Watson (1999) and Orphanides and van Norden (2003)).

(3) Predicting the monthly change in the U.S. unemployment rate with an AR(12) model (Montgomery, et al. (1998) and Terui and van Dijk (2002)).

(4) 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 (Mankiw and Miron (1986) and Lange, Sack, and Whitesell (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 (Paye and Timmermann (2002), Pesaran and Timmermann (2002a), and Rapach and Wohar (2002)).

(6) Predicting the monthly change in the U.S. dollar-Swiss franc spot exchange rate with interest differentials at 1, 3, 6, and 12 months (Clarida and Taylor (1997) and Clarida, et al. (2003)).

(7-12) Predicting quarterly GDP growth in the non-U.S. G7 countries (henceforth, G6 countries) with AR models (Min and Zellner (1993) and Stock and Watson (2004)).

In this empirical analysis, we consider the same forecast methods included in the Monte Carlo analysis, with some minor modifications.[11] Rather than allowing a range of arbitrary rolling window sizes, we examine forecasts based on just a 10-year window (with the exception of the exchange rate application, for which data limitations lead us to shorten the window to six years). We also, by necessity, drop consideration of the rolling forecast based on the known R_t^* and the shrinkage forecast using the known break date. Finally, in the break analysis, we impose a minimum break segment length of five years of data – 20 quarterly observations or 60 monthly observations (with the exchange rate sample relatively short, in that case we shorten the minimum segment length to 36 observations).

5.2 Results

In a broad sense, the application results presented in Tables 6 and 7 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 3-month interest rate-term spread for 1990-03) but inferior forecasts

[11] Note also that our Andrews (1993) tests use Wald statistics with heteroskedasticity-robust variances.

when the timing is not (as in the same application results for 1976-89). Here, too, it seems, the methods that are capable of performing the best when a break may have occurred tend to perform the worst when the model has been stable. The reverse CUSUM method provides a perhaps even more stark example of this pattern. The CUSUM method can produce nearly the most accurate forecast (3-month interest rate-term spread application, 1990-03) but often produces one of the worst (same application, 1976-89).

Such broad similarities aside, one particularly notable result is the difficulty of beating the recursive approach.[12] Despite the extant evidence of instability in many of the applications considered, the recursive forecast is frequently the best. Perhaps most strikingly, in the core inflation and stock return applications, none of the alternative approaches yields a forecast RMSE materially smaller than the recursive, for either of the reported sample periods. Indeed, in several cases, the alternative forecasts have RMSEs roughly 20 percent larger than the recursive forecast. The same basic pattern applies in the exchange rate and German GDP examples, although the failures relative to the recursive are not as large as in the core inflation and stock return cases. That said, the recursive approach does not seem to perform as strongly in the G6 GDP growth applications in Table 7 as in the U.S. applications in Table 6.

Despite the general difficulty of improving on the recursive method, there are some approaches that, in terms of RMSE, usually forecast as well or better. 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

---

[12] 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.

estimates performs very well, and perhaps 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. When the recursive forecast is best, the Bayesian shrinkage of 10 year estimates essentially matches the recursive RMSE.

Consider, for example, the 3-month interest rate-term spread application. For 1976-89 the 10-year shrinkage forecast is essentially as accurate as the top-ranked recursive forecast, with a RMSE ratio of 1.007; a rolling forecast based on 40 observations has a RMSE ratio of 1.027. For 1990-03, the 10-year shrinkage forecast's RMSE ratio is .930, compared to the best RMSE ratio of .747 provided by a simple rolling forecast. A shrinkage forecast that uses a post-break sample without requiring the estimated break to be significant doesn't perform as well: the *shrinkage: post-break R (all)* RMSE ratios are 1.010 and 1.028 for 1976-89 and 1990-03, respectively. Bayesian model averaging also doesn't perform as well, yielding RMSE ratios of 1.007 and .967 when a large prior weight is placed on the recursive model (the forecast based on BMA with equal weights has RMSE ratios of 1.034 and .911). In this application, discounted least squares performs as well as shrinkage of rolling estimates based on 10 years of data, with RMSE ratios of 1.009 and .926 for 1976-89 and 1990-03, respectively.

In the GDP-interest rates application, the 10-year shrinkage forecast improves on the accuracy of the recursive forecast in both periods, with a RMSE ratio of .981 for 1976-89 and .946 for 1990-03, essentially matching the performance of the forecast based on a 10 year rolling window. The shrinkage forecast that uses a post-break window without requiring the estimated break to be significant (the *shrinkage: post-break R (all)* forecast) yields RMSE ratios of .988 and .993 for 1976-89 and 1990-03, respectively. In this application, Bayesian model averaging performs roughly as well as 10-year shrinkage. For instance, for 1990-03, BMA with equal prior weight on all models yields an RMSE ratio of .959; BMA with a large prior weight on the recursive yields an RMSE ratio of .980. Discounted least squares also performs well, with RMSE ratios of .980 for 1976-89 and .930 for 1990-03. 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 comes 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, predictions based on either the optimal combination of recursive and post-break forecasts or a rolling window of an estimated R_t^* (α=0) observations are sometimes at least as accurate as recursive forecasts (as in the GDP-interest rates and Japan GDP applications), but sometimes significantly less accurate (as with the stock return and 3-month interest rate-term spread examples).

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.

<u>Assumption 1:</u> (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.

<u>Assumption 2</u>: The parameters are estimated using OLS.

<u>Assumption 3:</u> (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).

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

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

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

$$+ \int_1^{1+\lambda_P} [(1 - \alpha^2(s))s^{-2}(\int_0^s g(r)dr)' B^{-1}(\int_0^s g(r)dr)$$

$$- (1 - \alpha(s))^2 \lambda_R^{-2}(\int_{s-\lambda_R(s)}^s g(r)dr)' B^{-1}(\int_{s-\lambda_R(s)}^s g(r)dr)]ds$$

$$- 2\int_1^{1+\lambda_P} \alpha(s)(1 - \alpha(s))s^{-1}\lambda_R^{-1}(s)(\int_0^s g(r)dr)' B^{-1}(\int_{s-\lambda_R(s)}^s g(r)dr)ds \}$$

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

Appendix 2: Application Details

Unless otherwise noted, all data are taken from the FAME database of the Board of Governors. All data end in 2003:Q4 or December 2003. Growth rates and inflation rates are calculated as log changes. 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. In all cases the forecasting model includes a constant in the set of predictors.

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. Core CPI inflation	core CPI inflation (qly)	four lags (AR(4))	start point: 1958:Q2 data for 1967-83 are the BLS' housing-consistent series instead of the published core CPI
3. Unemployment rate	Δ unemployment rate (mly)	12 lags (AR(12))	start point: 1954:1
4. 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
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: January 1953 excess return = return less 1-month interest rate d-p ratio based on average of dividends from *t-11* to *t* S&P 500 dividend data from Global Insight; 1-month interest rate from Kenneth French's website
6. U.S.-Switz. exchange rate	U.S.-Switzerland ex. rate (nominal, mly, end of month)	two lags of: the change in the spot rate and the U.S.-Switz. differential in 1, 3, 6, and 12 month interest rates (all end of month)	start point: August 1973 interest rates from Global Insight
7. G6 GDP growth	growth rate of real GDP (qly) in the non-U.S. G7 countries[13]	lags of growth (AR model), order determined with AIC	start points: 1961:3 Canada 1964:2 France 1961:2 Germany 1960:4 Italy 1956:2 Japan 1955:3 UK

[13] We smoothed some outlier observations due to factors such as strikes and German reunification as follows. (1) The growth rate of German GDP in 1991:Q1 was set to the forecast from an AR(4) model fit with data from 1961 through 1990:Q4. (2) Output in France in 1968:Q2 was calculated by interpolating between 1968:Q1 and 1968:Q3.

References

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

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.

Berger, J. and Pericchi, L. (2004). "Training samples in objective Bayesian model selection," Annals of Statistics, 18, forthcoming.

Cho, I.K., N. Williams, and T.J. Sargent (2002): "Escaping Nash Inflation," Review of Economic Studies, 69, 1-40.

Clarida, R.H., L. Sarno, M.P. Taylor, and G. Valente (2003): "The Out-of-Sample Success of Term Structure Models as Exchange Rate Predictors: a Step Beyond," Journal of International Economics, 60, 61-83.

Clarida, R.H. and M.P. Taylor (1997): "The Term Structure of Forward Exchange Premiums and the Forecastability of Spot Exchange Rates: Correcting the Errors," Review of Economics and Statistics, 79, 353-61.

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.

de Jong, R.M. and Davidson, J. (2000): "The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals I: Weakly Dependent Processes," Econometric Theory, 16, 621-42

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. Timmermann (2004): "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions," Journal of Econometrics, 122, 47-79.

Engle, R. and A.D. Smith, (1999) "Stochastic Permanent Breaks," Review of Economics and Statistics, 81, 553-574.

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

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. (1997): “Approximate Asymptotic P Values for Structural-Change Models,” Journal of Business and Economic Statistics, 15, 60-67.

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.

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.

Maheu, J.M. and S. Gordon (2004): “Learning, Forecasting and Structural Breaks”, manuscript, University of Toronto.

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.

Montgomery, A.L., V. Zarnowitz, R. Tsay, and G.C. Tiao (1998): “Forecasting the U.S. Unemployment Rate,” Journal of the American Statistical Association, 93, 478-93.

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, 9, 495-510.

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

Pesaran, M.H., D. Pettenuzzo and A. Timmermann (2004): "Bayesian Regime Averaging for Time Series subject to Structural Breaks," 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 (2003): "Forecasting Output and Inflation: The Role of Asset Prices," Journal of Economic Literature, 41, 788-829.

Stock, J.H. and M.W. Watson (2004): "Combination Forecasts of Output Growth in a Seven–Country Data Set", Journal of Forecasting, 23, 405-30.

Terui, N., and H.K. van Dijk, (2002): "Combined Forecasts from Linear and Nonlinear Time Series Models," International Journal of Forecasting, 18, 421-38.

Wang, J. and E. Zivot (2000): "A Bayesian Time Series Model of Multiple Structural Changes in Level, Trend, and Variance," Journal of Business and Economic Statistics, 18, 374-386

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	explanation
recursive	coefficient estimates based on all available data
rolling: R=20	coefficient estimates based on 20 most recent observations
rolling: R=40	same as above, except that R = 40
shrinkage: R=20	coefficient estimates based on 20 most recent observations, with shrinkage toward recursive estimates, using (6)
shrinkage: R=40	same as above, except that R = 40
rolling: reverse CUSUM R	coefficient estimates based on data since break identified by reverse order CUSUM (1% sig.level)
rolling: sup Wald R	coefficient estimates based on data since break identified by Andrews' (1993) sup Wald test for a single break (5% sig.level)
rolling: known R* (α=0)	coefficient estimates based on 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)	coefficient estimates based on 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 break date	coefficient estimates based on post-break window, using the known break date imposed in the DGP, with shrinkage toward recursive estimates, using (6)
shrinkage: sup Wald R (5%)	coefficient estimates based on post-break window, using break dates identified as significant at the 5% level, with shrinkage toward recursive m estimates, using (6)
shrinkage: sup Wald R (all)	coefficient estimates based on post-break window, using least squares estimate of break date regardless of test significance, with shrinkage toward recursive estimates, using (6)
opt. combination: known	combination of the recursive forecast and a forecast based on rolling parameter estimates from the post-break period, with weights determined using (5) and the known features of the DGP
opt. combination: estimated	combination of the recursive forecast and a forecast based on rolling parameter estimates from the post-break period, with weights estimated using (5), based on the results of the Andrews (1993) test (5% sig.level) and the estimated date of the break
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
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
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
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 break date	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
shrinkage: sup Wald R (5%)	1.000	1.000	1.000	1.000	1.001	1.001	1.001	1.001
shrinkage: sup Wald R (all)	1.005	1.005	1.005	1.005	1.003	1.004	1.003	1.003
opt. combination: known	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
opt. combination: estimated	1.008	1.011	1.011	1.010	1.028	1.026	1.025	1.023
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. The forecast approaches are defined in Table 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 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
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
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
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 break date	0.947	0.944	0.944	0.947	0.951	0.947	0.947	0.948
shrinkage: sup Wald R (5%)	0.965	0.958	0.956	0.955	0.963	0.956	0.953	0.951
shrinkage: sup Wald R (all)	0.947	0.943	0.944	0.947	0.951	0.947	0.946	0.947
opt. combination: known	0.854	0.860	0.871	0.887	0.843	0.838	0.842	0.853
opt. combination: estimated	0.941	0.928	0.926	0.928	0.924	0.902	0.892	0.888
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
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
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
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 break date	0.946	0.948	0.951	0.956	0.948	0.948	0.949	0.952
shrinkage: sup Wald R (5%)	0.958	0.957	0.958	0.962	0.953	0.952	0.952	0.954
shrinkage: sup Wald R (all)	0.946	0.947	0.951	0.956	0.947	0.947	0.948	0.951
opt. combination: known	0.890	0.898	0.908	0.922	0.866	0.870	0.875	0.887
opt. combination: estimated	0.936	0.935	0.939	0.946	0.906	0.902	0.901	0.906
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. The forecast approaches are defined in Table 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 $\lambda_B 100$.
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
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
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
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 break date	0.953	0.953	0.956	0.961	0.951	0.951	0.953	0.958
shrinkage: sup Wald R (5%)	0.975	0.972	0.972	0.973	0.973	0.969	0.968	0.969
shrinkage: sup Wald R (all)	0.955	0.954	0.957	0.962	0.955	0.954	0.955	0.960
opt. combination: known	0.905	0.911	0.920	0.934	0.919	0.917	0.923	0.934
opt. combination: estimated	0.985	0.977	0.977	0.978	1.006	0.992	0.990	0.987
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

Notes:

1. DGPs DGP 1-B2, DGP 2-B2, DGP 1-B1, and DGP 2-B1 are defined in Section 4.1. The forecast approaches are defined in Table 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 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
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
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
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 break date	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
shrinkage: sup Wald R (5%)	0.863	0.053	0.795	0.067	0.751	0.079	0.675	0.084
shrinkage: sup Wald R (all)	0.000	0.422	0.000	0.393	0.000	0.360	0.000	0.332
opt. combination: known	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
opt. combination: estimated	0.863	0.035	0.795	0.033	0.751	0.021	0.675	0.023
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, $\boldsymbol{\lambda_B}$ = .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
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
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
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 break date	0.000	0.923	0.000	0.974	0.000	0.989	0.000	0.995
shrinkage: sup Wald R (5%)	0.253	0.619	0.106	0.779	0.063	0.847	0.025	0.920
shrinkage: sup Wald R (all)	0.000	0.889	0.000	0.943	0.000	0.969	0.000	0.986
opt. combination: known	0.000	0.813	0.000	0.902	0.000	0.928	0.000	0.964
opt. combination: estimated	0.253	0.477	0.106	0.629	0.063	0.700	0.025	0.784
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. The forecast approaches are defined in Table 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 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: Results for U.S. Forecasting Applications

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

	GDP-interest rates		Core CPI Inflation	
	1976-89	*1990-03*	*1976-89*	*1990-03*
recursive	3.663	2.497	2.063	0.558
rolling: fixed R	0.978	0.959	1.048	1.102
shrinkage: fixed R	0.981	0.946	1.009	0.987
rolling: reverse CUSUM R	0.989	1.080	1.029	0.966
rolling: sup Wald R	0.991	0.991	1.254	1.168
rolling: estimated R* (α=0)	0.997	0.991	1.080	1.001
shrinkage: sup Wald R (5%)	0.992	0.993	1.029	1.013
shrinkage: sup Wald R (all)	0.988	0.993	1.029	1.013
opt. combination: estimated	0.991	0.991	1.150	1.116
BMA, equal prior prob.	0.976	0.959	1.049	1.053
BMA, large prior prob.	0.985	0.980	1.018	1.015
DLS	0.980	0.930	1.036	0.983
	Unemployment rate		**3-mo. Int. rate-spread**	
	1976-89	*1990-03*	*1976-89*	*1990-03*
recursive	0.192	0.142	1.623	0.543
rolling: fixed R	1.030	1.012	1.027	0.747
shrinkage: fixed R	1.000	0.994	1.007	0.930
rolling: reverse CUSUM R	1.029	1.115	1.083	0.807
rolling: sup Wald R	1.001	0.992	1.107	1.019
rolling: estimated R* (α=0)	1.000	1.000	1.061	1.024
shrinkage: sup Wald R (5%)	1.001	0.993	1.006	0.998
shrinkage: sup Wald R (all)	0.990	0.993	1.010	1.028
opt. combination: estimated	1.001	0.992	1.063	1.007
BMA, equal prior prob.	1.013	0.985	1.034	0.911
BMA, large prior prob.	0.999	0.991	1.007	0.967
DLS	1.065	1.025	1.009	0.926
	Stock returns		**U.S.-Switz. exchange rate**	
	1976-89	*1990-03*	*1980-89*	*1990-03*
recursive	4.558	4.386	4.095	3.195
rolling: fixed R	1.039	1.052	1.048	1.082
shrinkage: fixed R	1.004	1.001	1.005	0.996
rolling: reverse CUSUM R	1.022	1.024	1.090	1.143
rolling: sup Wald R	1.061	1.060	1.043	1.048
rolling: estimated R* (α=0)	1.057	1.057	1.019	1.040
shrinkage: sup Wald R (5%)	1.002	0.999	0.998	1.004
shrinkage: sup Wald R (all)	1.004	0.999	1.012	1.004
opt. combination: estimated	1.052	1.054	1.021	1.038
BMA, equal prior prob.	1.029	1.007	1.086	1.016
BMA, large prior prob.	1.004	1.001	1.018	1.000
DLS	1.028	1.015	1.028	1.002

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. Note that, for the fixed R rolling and shrinkage forecasts, R = 40 for the (quarterly) GDP, core inflation, and 3-month interest rate applications. R = 120 for the (monthly) unemployment and stock return examples and 72 for the (monthly) exchange rate application.
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.

Table 7: Results for G6 GDP Forecasting Applications

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

	Canada (AR(1))		France (AR(4))	
	1976-89	*1990-03*	*1976-89*	*1990-03*
recursive	3.547	2.381	2.087	1.897
rolling: R=40	0.989	0.863	1.009	1.034
shrinkage: R=40	0.984	0.941	0.981	0.993
rolling: reverse CUSUM R	0.960	0.982	1.017	1.159
rolling: sup Wald R	1.005	0.902	0.989	0.957
rolling: estimated R* (α=0)	1.002	0.906	0.991	0.957
shrinkage: sup Wald R (5%)	0.995	0.948	0.966	0.974
shrinkage: sup Wald R (all)	0.969	0.931	0.966	0.974
opt. combination: estimated	1.000	0.908	0.983	0.957
BMA, equal prior prob.	0.951	0.882	0.979	0.967
BMA, large prior prob.	0.977	0.935	0.973	0.980
DLS	0.962	0.904	0.975	0.987
	Germany (AR(4))		**Italy (AR(1))**	
	1976-89	*1990-03*	*1976-89*	*1990-03*
recursive	4.847	3.379	3.028	2.947
rolling: R=40	1.012	1.043	0.994	0.912
shrinkage: R=40	0.995	1.000	0.996	0.979
rolling: reverse CUSUM R	1.031	1.136	0.978	0.963
rolling: sup Wald R	1.025	0.991	1.004	0.932
rolling: estimated R* (α=0)	1.020	0.992	1.005	0.936
shrinkage: sup Wald R (5%)	1.008	0.994	1.001	0.972
shrinkage: sup Wald R (all)	1.008	0.994	0.996	0.972
opt. combination: estimated	1.018	0.992	1.003	0.932
BMA, equal prior prob.	1.031	0.990	0.997	0.925
BMA, large prior prob.	1.005	0.993	0.995	0.966
DLS	1.008	0.981	0.998	0.948
	Japan (AR(3))		**UK (AR(1))**	
	1976-89	*1990-03*	*1976-89*	*1990-03*
recursive	3.432	3.450	4.162	2.092
rolling: R=40	0.937	0.927	1.039	0.880
shrinkage: R=40	0.966	0.973	1.007	0.990
rolling: reverse CUSUM R	1.009	0.973	1.008	0.968
rolling: sup Wald R	0.917	0.944	1.002	0.814
rolling: estimated R* (α=0)	0.913	0.927	1.035	0.931
shrinkage: sup Wald R (5%)	0.950	0.959	1.000	0.958
shrinkage: sup Wald R (all)	0.950	0.959	0.998	0.958
opt. combination: estimated	0.905	0.944	0.998	0.825
BMA, equal prior prob.	0.928	0.932	1.017	0.925
BMA, large prior prob.	0.960	0.963	1.004	0.974
DLS	0.954	0.930	1.010	0.951

Notes:

1. The orders of the AR models for each country, determined with the AIC, are provided in the headers. Other details of the 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.