ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

WORKING PAPER SERIES

Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts

Authors	Todd E. Clark, and Michael W. McCracken
Working Paper Number	2008-028A
Creation Date	August 2008
Citable Link	https://doi.org/10.20955/wp.2008.028
Suggested Citation	Clark, T.E., McCracken, M.W., 2008; Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts, Federal Reserve Bank of St. Louis Working Paper 2008-028. URL https://doi.org/10.20955/wp.2008.028

Federal Reserve Bank of St. Louis, Research Division, P.O. Box 442, St. Louis, MO 63166

Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts*

Running head: Recursive and Rolling Forecasts

Todd E. Clark
Federal Reserve Bank of Kansas City
and
Michael W. McCracken [1]
Board of Governors of the Federal Reserve System

Abstract

This paper presents analytical, Monte Carlo, and empirical evidence on combining recursive and rolling forecasts when linear predictive models are subject to structural change. Using 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, we derive optimal observation windows and combining weights designed to minimize mean square forecast error. Monte Carlo experiments and several empirical examples indicate that combination can often provide improvements in forecast accuracy relative to forecasts made using the recursive scheme or the rolling scheme with a fixed window width.

Keywords: structural breaks, forecasting, model averaging.
JEL Nos.: C53, C12, C52

---

* Manuscript received April 2006; revised September 2007.

[1] We 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. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City, Board of Governors, Federal Reserve System, or any of its staff.

1. Introduction

In a universe characterized by 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 tends to lead to patterns in the decisions on whether or not to use all available data when constructing forecasts. The finance literature tends to construct forecasts using only a rolling window of the most recent observations. In the macroeconomics literature, while usage of rolling schemes seems to be increasing, it has historically been 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 2006), one reason for the rolling approach to be historically more common in finance than in macroeconomics may simply be that financial series are often substantially longer.[2]

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),

[2] See Fama and MacBeth (1973) for an early example of rolling windows in finance.

Stock and Watson (2003), Maheu and Gordon (2007), Koop and Potter (2004), and Pesaran, Pettenuzzo and Timmermann (2006) 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 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 jointly optimal observation window and combining weights. The optimal forecast combination rule we develop can be interpreted as a frequentist approach to shrinkage.

Of course, conventional Bayesian methods provide an alternative approach to shrinkage. For example, a Bayesian could place prior distributions on the pre-break coefficients and on the size of coefficient change at the possible break point. With the break point unknown, a range of models allowing different break points could then be averaged on the basis of posterior probabilities. We consider such alternatives in our Monte Carlo and empirical analyses.

The results in the paper suggest a benefit to combining recursive and rolling forecasts. In the theory, we show a weighted average forecast to be at least as accurate, and often more accurate, than any forecast based on a single estimation sample, even when the single sample is optimized to maximize forecast accuracy. In our Monte Carlo and empirical results, our proposed combination method consistently improves forecast accuracy. Moreover, in terms of forecast accuracy, our proposed method is at least competitive with the Bayesian alternatives we consider.

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.[3]

Finally, our work on the optimal choice of observation window extends recent work by Pesaran and Timmermann (2007). 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 length of the post-break period.

Our results, however, differ from those in Pesaran and Timmermann along several dimensions. First, we model the breakpoint process as local-to-zero rather than using direct, finite-sample magnitudes. By doing so we emphasize the importance of the choice of observation window in situations where structural break tests have little power. Second, by using our asymptotic approach we are able to obtain closed form solutions for the optimal window size in the presence of conditional heteroskedasticity and serial correlation in the regression error terms. Finally, while Pesaran and Timmermann's Monte Carlo analysis includes model combination — with models differing by the unknown date of the putative structural change — as a competitor to the optimal choice of observation window, we explicitly derive closed form solutions for the optimal combining weights.

[3] 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.

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 develops the optimal combination forecast. In section 4 we present Monte Carlo evidence on the finite sample effectiveness of combination, along with some Bayesian alternatives. Section 5 compares the effectiveness of the forecast methods in a range of empirical applications. The final section concludes. Details pertaining to theory are presented in an appendix.

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 observation window. A detailed set of technical assumptions, sufficient for the results, are given in the appendix. 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. Note that, for simplicity, we use the term "rolling" to refer to model estimates and forecasts that, in our theoretical results, are based on a partial sample of the data. In common practice, rolling estimation uses a fixed sample size; in our results, the size of the partial sample is allowed to change as forecasting moves forward in time.

For tractability, our theoretical results are based on a single, discrete, structural break, modeled as a local rather than global break. In practice, to be sure, some research suggests the importance of multiple or stochastic breaks (e.g., Pesaran and Timmermann (2002) and Rapach and Wohar (2006)). However, there are enough studies finding just a single break (e.g., Hooker (2002) and

Estrella, Rodrigues, and Schich (2003)) to suggest practical value for our theoretical results. The local approximation, used in much of the literature on structural break tests, makes the analytics more tractable (as noted above, the local asymptotics allow us to derive closed form solutions under assumptions somewhat more general than in related prior work), and is consistent with the common view that, in practice, breaks are small enough that conventional tests have low power (see, e.g., the power discussion in Cogley and Sargent (2005)). Of course, large breaks will have different theoretical implications (see Inoue and Kilian, 2003, pp. 22-24). The empirical applications considered in section 5 will shed light on the practical value of our analytical results based on a single break and local asymptotics.

2.1 Environment

The possibility of structural change is modeled using a sequence of linear DGPs of the form[4]

(1) $$y_{T,t+\tau} = x'_{T,t}\beta^*_{T,t} + u_{T,t+\tau} \qquad \beta^*_{T,t} = \beta^* + T^{-1/2}1(t \geq T_B)\Delta\beta$$
$$Ex_{T,t}u_{T,t+\tau} \equiv Eh_{T,t+\tau} = 0 \text{ for all } t.$$

In this formulation, at time T_B (modeled as a fixed proportion of the initial forecast origin $[\lambda_B T]$) there is structural change in the regression parameter vector β^* of magnitude $T^{-1/2}\Delta\beta$. Note that we allow the τ-step ahead predictand $y_{T,t+\tau}$, the predictors $x_{T,t}$ and the error term $u_{T,t+\tau}$ to depend upon T. 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.

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

At each forecast origin $t = T,...T+P$, where P denotes the number of forecasts, 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 which may include lagged dependent variables. τ-step ahead 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 $\hat{\beta}_{R,t} = \arg\min_{\beta} t^{-1}\sum_{s=1}^{t-\tau}(y_{s+\tau} - x_s^{'}\beta)^2$ and $\hat{\beta}_{L,t} = \arg\min_{\beta} R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}(y_{s+\tau} - x_s^{'}\beta)^2$ 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 $\hat{u}_{L,t+\tau}^2 = (y_{t+\tau} - x_t^{'}\hat{\beta}_{L,t})^2$.

As detailed in the appendix, in deriving our theoretical results we maintain that the DGP is a linear regression subject to local structural change. The structural change is nonstochastic and of a small enough magnitude that the observables are asymptotically mean square stationary.[5] Despite various technical conditions—sufficient to insure that certain partial sums of $h_{T,t+\tau} \equiv x_{T,t}u_{T,t+\tau}$ converge weakly to standard Brownian motion—we allow the model errors to form a conditionally heteroskedastic MA$(\tau - 1)$ process.

Finally, in our derivations we generalize assumptions made in West (1996) that require the length of the observation window (associated with the rolling scheme) to be fixed so that

[5] Loosely speaking, an array $x_{T,t}$ is asymptotically mean square stationary if in large samples it is weakly stationary. As an example consider the AR(1) process $y_{T,t} = (\alpha + T^{-1/2}1(t \geq T_B)\Delta\alpha) + \beta y_{T,t-1} + u_{T,t}$ with structural change in the intercept. For $t < T_B$ $Ey_{T,t} = \alpha/1-\beta$ and for $t \geq T_B$ $Ey_{T,t} = (\alpha + T^{-1/2}\Delta\alpha)/1-\beta$. While it is true that the structural change implies that $y_{T,t}$ is nonstationary in finite samples, in large samples such nonstationarities vanish. See Hansen (2000) for a more rigorous definition of asymptotic mean square stationarity.

$\lim_{T\to\infty} R_t / T = \lambda_R \in (0,1)$. Instead, we weaken that assumption so that $R_t / T \Rightarrow \lambda_R(s) \in (0,s]$, $1 \le s \le 1+\lambda_P$ (where $\lim_{T\to\infty} P / T = \lambda_P \in (0,\infty)$) and hence the observation window is allowed to change with time as evidence of instability is discovered.

2.2 Theoretical results on the tradeoff

Our approach to understanding the bias-variance tradeoff is based upon an analysis of $\sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau} - \hat{u}^2_{L,t+\tau})$, the difference in the (normalized) MSEs of the recursive and rolling forecasts.[6] As detailed in Theorem 1 in the appendix, we show that this statistic has an asymptotic distribution that can be decomposed into three terms. 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. From that decomposition, we are able to establish that the bias-variance tradeoff depends on factors such as the size of the rolling window and the size of the coefficient break. However, providing a complete analysis of the distribution of the relative accuracy measure is difficult because we do not have a closed form solution for its density. Therefore, we proceed in the remainder of this section to focus on the mean (rather than the distribution) of the bias-variance tradeoff when there are either no breaks or a single break.[7]

---

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

[7] By taking this approach we are using the fact that under our assumptions, notably the L^2-boundedness portion of Assumption 3, $\sum_{t=T}^{T+P}(\hat{u}^2_{R,t+\tau} - \hat{u}^2_{L,t+\tau})$ is uniformly integrable and hence the expectation of its limit is equal to the limit of its expectation.

The results presented below use the following additional notation: V denotes the long-run variance of the OLS orthogonality vector $h_{t+\tau}$, and B denotes the second moments of the predictors, specifically, $B = \lim_{T\to\infty}(Ex_{T,t}x_{T,t}^{'})^{-1}$.

2.3 The case of no break

We can precisely characterize the asymptotic mean of $\sum_{t=T}^{T+P}(\hat{u}_{R,t+\tau}^2 - \hat{u}_{L,t+\tau}^2)$ in the case of no breaks. Using the representation from Theorem 3.1 in the appendix we obtain

(2) $$E[\lim_{T,P\to\infty}\sum_{t=T}^{T+P}(\hat{u}_{R,t+\tau}^2 - \hat{u}_{L,t+\tau}^2)] = tr(BV)\int_1^{1+\lambda_P}(\frac{1}{s} - \frac{1}{\lambda_R(s)})ds$$

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

Now suppose that a permanent structural change, of magnitude $T^{-1/2}\Delta\beta \neq 0$, 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 the appendix 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

$$E[\lim_{T,P\to\infty} \sum_{t=T}^{T+P} (\hat{u}^2_{R,t+\tau} - \hat{u}^2_{L,t+\tau})] = \int_1^{1+\lambda_P} [tr(BV)(\frac{1}{s} - \frac{1}{\lambda_R(s)})]ds \; +$$

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

(b) If $\lambda_R(s) \leq s - \lambda_B$ for all $s \in [1, 1+\lambda_P]$ then $E[\lim_{T,P\to\infty} \sum_{t=T}^{T+P} (\hat{u}^2_{R,t+\tau} - \hat{u}^2_{L,t+\tau})] =$

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

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—optimal for forecasting in the presence of a single structural change in the regression coefficients—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

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

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

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

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. In fact, as the break magnitudes ($\Delta\beta' B^{-1}\Delta\beta$) become large, or the precision ($tr(BV)$) of the parameter estimates shrinks to zero, we obtain the intuitive result that the observation window includes only post-break data.

Note, however, that the term $\Delta\beta' B^{-1}\Delta\beta$ is a function of the local-to-zero break magnitude $\Delta\beta$ and 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 we let $\Delta\hat{\beta}$ and $\hat{T}_B$ denote OLS estimates of the break magnitude ($T^{-1/2}\Delta\beta$) 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.[9]

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

One final note on the formulae in Corollary 2.2 and (4). In Corollary 2.2, we use break magnitudes that are 'local-to-zero' 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. Moreover, as noted in Elliott (2005), it captures the practical reality of forecasting in an environment in which it is difficult to detect structural change.

Unfortunately, by taking this approach we arrive at formulas that depend upon unknown, local-to-zero break magnitudes that cannot be consistently estimated (Bai (1997)). Regardless, we continue to use OLS estimates of these break magnitudes and dates to estimate (inconsistently) the implied optimal observation window. Our Monte Carlo experiments indicate that the primary difficulty is not the inconsistency of our estimate of the optimal observation window; rather, the

[9] 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 our Monte Carlo and empirical implementations, $tr(BV)$ is estimated as $tr(BV) = tr[(t^{-1}\sum_{j=1}^{t} x_j x_j')^{-1}(t^{-1}\sum_{j=1}^{t} \hat{u}_{j+\tau}^2 x_j x_j')]$, where $\hat{u}$ refers to the residuals from estimates of the forecasting model using data from 1 to t.

primary difficulty is break identification and dating.[10] 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. 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.

3.1 Optimal combination

With linear models, a linear combination of the recursive and rolling 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 the appendix, 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. Note again that the

[10] Elliott (2005) reaches a similar conclusion in the context of forecasting.

windows are pointwise optimal because they are derived by maximizing the components of the expected loss differential over the duration of forecasting.

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

$$E[\lim_{T,P\to\infty} \sum_{t=T}^{T+P} (\hat{u}^2_{R,t+\tau} - \hat{u}^2_{W,t+\tau})] = tr(BV)\int_1^{1+\lambda_P} (1 - \alpha(s))^2 (\frac{1}{s} - \frac{1}{\lambda_R(s)})ds +$$

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

(b) If $\lambda_R(s) \leq s - \lambda_B$ for all $s \in [1, 1+\lambda_P]$ then $E[\lim_{T,P\to\infty} \sum_{t=T}^{T+P} (\hat{u}^2_{R,t+\tau} - \hat{u}^2_{W,t+\tau})] =$

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

Corollary 3.2: In the presence of a single break in the regression parameter vector, the pointwise (jointly) optimal window width and combining weights satisfy $(\lambda_R^*(s), \alpha^*(s)) =$

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

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, under the assumptions on the breakpoint process considered here, 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, but to focus instead on forecast combination. Corollary 3.2 therefore provides a formal justification for the model averaging Pesaran and Timmermann (2007) include in their Monte Carlo analysis. While our formal results only apply in our single break setup, the intuition that applies in the single break case should also go through in alternative setups, such as a case of multiple breaks. Accordingly, the basic finding that combination is optimal should, under similar data features, extend to other cases such as multiple breaks. However, it would be very difficult to prove analytically, and we have not done so in this paper.

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.

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

4. Monte Carlo Results

We use Monte Carlo simulations of multivariate data-generating processes to evaluate, in finite samples, the performance of the forecast methods described above. In these experiments, the DGP relates a scalar 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 Bayesian methods described below. Performance is evaluated on the basis of average MSEs across Monte Carlo draws.

4.1 Experiment design

To ensure the practical relevance of our results, we use two DGPs based on relationships estimated from quarterly U.S. data, taken from variants of applications we consider in the next section. We base DGP 1 on the relationships among GDP growth (y), the spread between the 10-year Treasury bond and 3-month Treasury bill rates (x_1), and the change in the 3-month Treasury bill rate (x_2):[12]

$$\begin{aligned}
&y_t = (.2 + d_{t-1}\Delta b_y) y_{t-1} + (2.0 + d_{t-1}\Delta b_{x1}) x_{1,t-1} + (1.4 + d_{t-1}\Delta b_{x2}) x_{2,t-1} + u_t \\
&x_{1,t} = 1.1 x_{1,t-1} - .2 x_{1,t-2} + v_{1,t} \\
&x_{2,t} = .3 x_{2,t-1} - .3 x_{2,t-2} + v_{2,t} \\
&u_t, v_{1t}, v_{2t} \sim N(0, \Sigma) \\
&\Sigma = \begin{pmatrix} 10.4 & & \\ -.2 & .3 & \\ .6 & -.3 & .5 \end{pmatrix} \\
&d_t = 1(t \geq \lambda_B T)
\end{aligned}$$

In our baseline experiments, the size of the coefficient break is taken from the empirical estimates: $(\Delta b_y, \Delta b_{x1}, \Delta b_{x2}) = (0.0, -1.8, -1.0)$. We also consider experiments with a break half as large as in the baseline case and a break twice as large as in the baseline.

[12] We estimated the relationship with quarterly 1953-2006 data, imposing an Andrews (1993) test-identified break in 1984. The estimated relationships include intercepts, which we exclude from the DGP (but not the forecasting models) for simplicity.

We base DGP 2 on the relationships among the change in CPI inflation (y) and two common business cycle factors (x_1, x_2): [13]

$$
\begin{aligned}
& y_t = (-.5 + d_{t-1}\Delta b_{y1}) y_{t-1} + (-.3 + d_{t-1}\Delta b_{y1}) y_{t-2} + (.2 + d_{t-1}\Delta b_{x1}) x_{1,t-1} + (-.2 + d_{t-1}\Delta b_{x2}) x_{2,t-1} + u_t \\
& x_{1,t} = .8 x_{1,t-1} - .1 x_{1,t-2} + v_{1,t} \\
& x_{2,t} = .8 x_{2,t-1} + v_{2,t} \\
& u_t, v_{1t}, v_{2t} \sim N(0, \Sigma) \\
& \Sigma = \begin{pmatrix} 1.6 & & \\ 0.0 & 6.7 & \\ -.2 & -.7 & 2.4 \end{pmatrix} \\
& d_t = 1(t \geq \lambda_B T)
\end{aligned}
$$

In our baseline experiments, the size of the coefficient break is taken from the empirical estimates: $(\Delta b_y, \Delta b_{x1}, \Delta b_{x2}) = (-.1, -.2, .1, .2)$. We also consider experiments with a break half as large as in the baseline case and a break twice as large as in the baseline.

In each experiment, with post-war quarterly data in mind, we conduct 5000 simulations of data sets of 180 observations (not counting the initial observations necessitated by the lag structure of the DGP). The data are generated using innovation draws from the normal distribution and the autoregressive structure of the DGP.[14] We set T, the number of observations preceding the first forecast date, to 100, and consider forecast periods of various lengths: P = 1, 20, 40, and 80

[13] We estimated the relationship with quarterly (rather than monthly in the interest of keeping tractable the Monte Carlo time required) 1960-2006 data, imposing an Andrews (1993) test-identified break in 1980. The quarterly factor index values are within-quarter averages of monthly factors. For convenient scaling of the reported residual covariance matrix, the factors were multiplied by 10 prior to DGP estimation. The estimated relationships include intercepts, which we exclude from the DGP (but not the forecasting models) for simplicity.

[14] 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.

(corresponding to λ_P = .01, .2, .4, .6, and 1.0). For each value of P, forecasts are evaluated over the period $T+1$ through $T+P$.

We present results for experiments with two different break dates (a single break in each experiment), at observations 60 and 80 (corresponding to λ_B = .6 and .8).

4.2 Forecast approaches: combination and Bayesian model averaging

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

$$y_{t+1} = b'X_t + e_{t+1},$$

where $X_t = (1, y_t, x_{1,t}, x_{2,t})'$ for DGP 1 and $X_t = (1, y_t, y_{t-1}, x_{1,t}, x_{2,t})'$ for DGP 2. Table 1 details all of the forecast methods. As to the particulars of the implementation of our proposed forecasts, we note the following.

1. Our break tests are based on the full set of forecast model coefficients. For a data sample from observation 1 through the forecast origin t, we test for a break in the middle *t-40* observations (i.e., we impose a minimum segment length of 20 periods). The break test analysis is performed in real time, with tests applied at each forecast origin.
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.
3. Our results using break tests are based on the Andrews (1993) test for a single break, with a 2.5% significance level.[15] In results not reported in the interest of brevity, we considered

[15] 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 (2005) in the context of causality testing, repeated tests in such real time analyses with the

various alternatives, including the reverse order CUSUM method proposed by Pesaran and Timmermann (2002) and the BIC criterion of Yao (1988) and Bai and Perron (2003) (which allows for multiple breaks). While these approaches may have advantages in other settings, in our Monte Carlo experiments and empirical applications, the Andrews test approach generally performed better.

4. 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.

5. In light of the difficulty of identifying breaks in small samples and the potentially positive impact of forecast combination, we report results for an optimal combination of the recursive forecast with the fixed rolling window (40 observation) forecast. The combination weight is estimated using equation (5), assuming a break 10 years prior to the forecast origin. Admittedly, this 10-year window specification is somewhat arbitrary. With different break timing, the same window choice might not work as well. In practice, though, empirical forecast studies commonly use similar window sizes. Moreover, the 10-year window proves to work well in the applications in section 5.

Insert Table 1 here

The forecast methods for which we report results include Bayesian methods that might be considered natural alternatives to our proposed combination forecasts.[16] These Bayesian forecasts are based on a model that allows a single break in the coefficients sometime in the estimation

---

use of standard critical values will result in spurious break findings. However, in our context, in DGPs with breaks, performance could deteriorate, because not enough breaks would be found.

[16] Our Bayesian implementation is related to those of Wang and Zivot (2000) and Hultblad and Karlsson (2006).

sample (specifically, sometime in the middle *t-40* observations of a sample ending in *t*):

$y_{t+1} = x_t'(b + d_t \Delta b) + e_{t+1}, d_t = 1(t \geq break\ date)$. In the interest of presuming no break unless the data indicate otherwise, we use a loose prior for the pre-break coefficients (*b*) and allow a potentially informative prior for the coefficient shifts (Δb). We set the prior standard deviation on all *b* elements to be 1000σ and the standard deviation on all Δb elements to $\lambda\sigma$, where λ is a hyperparameter determining the tightness of the prior (all prior covariances are 0). All prior means are 0. For tractability, we use the textbook Normal-Gamma form for the prior, which yields a Normal-Gamma posterior.[17]

As to the hyperparameter setting, we consider two alternative approaches. First, in line with common BVAR practice (embedded, for example, in the defaults of Estima's RATS software), we fix λ at 0.2. Second, we consider a grid of values for λ, ranging from .0001 (which essentially corresponds to no break) to 1000 (which essentially results in a post-break rolling estimate), and use the λ value delivering the highest marginal likelihood.[18]

Of course, the break date needed in this Bayesian approach is not actually known. In results not reported, we considered (i) a fixed break date of 10 years prior to the forecast origin and (ii) the break date that delivers the highest marginal likelihood. However, in terms of point accuracy, forecasts based on a single model or break date were generally dominated by forecasts obtained by averaging across all possible break dates (in the middle *t-40* observations of the sample of *t*

---

[17] We use a loose prior for the inverse of the residual variance, of the form G(1/v,1), where v is set to .9 times the sample variance of the dependent variable estimated with data up to the forecast origin.

[18] The other hyperparameter values in the grid are .002, .1, .2, 4, 1.6, 4, 20, 100, and 400.

observations), with each possible break date/model/forecast weighted by its posterior probability.[19] We report one Bayesian model average forecast obtained with a fixed $\lambda = .2$ setting and another that (at each point in the forecast sample) uses the setting delivering the highest marginal likelihood.

4.3 Simulation results

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.

4.3.1 Average MSEs in baseline experiments

Table 2 reports results from our baseline Monte Carlo experiments, in which the sizes of the breaks in the DGPs match the estimates based on U.S. data. In these experiments, the forecasts based on the known features of the DGPs (break timing, size, and population moments) confirm the broad implications of the theoretical results in sections 2 and 3. Specifically, the combined forecast based on the known optimal α_t^* (*opt. comb.: known*) is more accurate than the forecast based on the known optimal estimation sample (*rolling: known R**), which is in turn more accurate than the forecast based on the known post-break estimation sample (*rolling: known break R*). And, in these experiments, the coefficient breaks are large enough that the forecast based on the post-break sample is more accurate than the forecast based on the full sample. For example, when the break in DGP 1 occurs at observation 80, the *rolling: known break R*, *rolling: known R**, and *opt. comb.: known* forecasts for the $P = 1$ sample have MSE ratios of, respectively, .979, .950, and .900.

Moreover, in line with our theory, the advantages of the optimal sample and combination forecasts over the post-break sample forecast tend to decline as the break moves further back in

---

[19] The posterior probability is calculated using the conventional Normal-Gamma analytical formula for the marginal likelihood.

time (relative to the forecast origin). In the experiments with a break at observation 80, differences in the accuracies of the three aforementioned forecasts are quite small for the P = 80 sample, at MSE ratios of .914, .912, and .898 in the DGP 1 results. Similarly, the advantages of the optimal sample and optimal combination forecasts over the post-break sample forecast are generally smaller in experiments with a break at observation 60 than in experiments with a break at observation 80. For example, when the break in DGP 1 occurs at observation 60, the *rolling: known break R*, *rolling: known R**, and *opt. comb.: known* forecasts for the P = 1 sample have MSE ratios of, respectively, .936, .933, and .915 (compared to .979, .950, and .900 when the break occurs at observation 80).

Insert Table 2 here

Not surprisingly, feasible forecasts based on estimates of the break date and size and other data moments are less accurate than the infeasible forecasts based on the known break date and size and moments. Nonetheless, the aforementioned implications of our theory continue to hold, although less dramatically than in the known moment case. In Table 2's baseline experiments, the estimated optimal combination forecast is slightly more accurate than the forecast based on the estimated optimal sample, which is in turn more accurate than the forecast based on the estimated post-break sample. For example, over the P = 20 sample, the *rolling: post-break R*, *rolling: estimated R**, and *opt. comb.: estimated* forecasts have MSE ratios of, respectively, .986, .964, and .962 in the DGP 1 experiment.

Much of the accuracy gap between the feasible methods of optimal sample and combination forecasting and the theoretical, infeasible methods appears to be attributable to difficulties in identifying whether a break occurred and when (difficulties perhaps not surprising in light of extant evidence of size and power problems with break tests). If we impose the known break date in determining the post-break sample and estimating the optimal sample size and combination weight (forecasts not reported in the tables in the interest of brevity), we obtain forecasts nearly as accurate as the *rolling: known break R*, *rolling: known R** and *opt. comb.: known* forecasts.

Accordingly, accuracy might be improved by simply imposing an arbitrary break date in model estimation and combination. Such an approach is not entirely uncommon; studies such as Swanson (1998) and Del Negro et al. (2007) have used rolling window sizes seemingly arbitrarily set, ranging from 10 to 30 years of data. We therefore consider two forecasts that suppose a break occurred 40 observations (10 years of quarterly data) prior to the end of the estimation sample/forecast origin: one based on a rolling estimation sample of 40 observations, and another obtained as an estimated optimal combination of the recursive forecast with the 40 observation rolling sample forecast.

In Table 2's results, imposing a break 40 observations prior to each forecast origin significantly improves the performance of our proposed optimal combination approach – enough that, among the feasible non-Bayesian forecasts in Table 2, the resulting optimal combination forecast is the most accurate. For example, with DGP 1, over the $P = 20$ sample, the *opt. comb.: fixed R* forecast has an MSE ratio of .914, compared to MSE ratios of .962 and .884 for the *opt. comb.: estimated* and *opt. comb.: known* forecasts. Admittedly, there are cases in Table 2, such as with DGP 1 and $P = 1$, in which the *opt. comb.: fixed R* forecast has little or no advantage over the forecast (*rolling: fixed R*) based on an arbitrary rolling sample of 40 observations. However, for larger forecast samples, the combination forecast is more accurate than the fixed rolling window forecast. For instance, with DGP 2 with a break at observation 60 and $P = 80$, the *opt. comb.: fixed R* forecast has an MSE ratio of .958, while the *rolling: fixed R* forecast yields an MSE ratio of .992. The improvement with larger samples reflects the fact that, as forecasting moves forward in time, more of the available data come to reflect the post-break sample, such that it is increasingly advantageous to incorporate information from the full sample of data (as the combination forecast does, by putting weight on the recursively estimated model) rather than just using the most recent data.

In Table 2's baseline experiments, our proposed optimal combination forecast based on a fixed break date of 40 observations prior to the forecast origin is competitive with Bayesian methods. For example, over the $P = 80$ forecast sample, the BMA forecast with the fixed prior variance hyperparameter yields MSE ratios of, respectively, .917 and .892 in DGPs 1 and 2 with a break at observation 80. The *opt. comb.: fixed R* forecast yields corresponding MSE ratios of .930 and .925. Of the two BMA forecasts, in the baseline experiments, imposing a fixed hyperparameter value of $\lambda = .2$ tends to yield forecasts slightly more accurate (more so for smaller forecast samples than larger forecast samples) than those obtained by choosing at each forecast origin the hyperparameter value that maximizes the marginal likelihood. Continuing with the same example, the BMA forecast with an optimized prior variance hyperparameter yields MSE ratios of .931 (DGP 1) and .898 (DGP 2).

4.3.2 Average MSEs in experiments with smaller and larger breaks

In broad terms, the results described above continue to hold in Monte Carlo simulations of DGPs in which the break in coefficients is half as small as or twice as large as the break imposed in the baseline simulations. For example, in the smaller break results reported in the upper panel of Table 3, the *opt. comb.: fixed R* is the most accurate of all of the feasible forecasts based on an estimated optimal estimation sample or combination. For instance, with a forecast sample of $P = 20$, this forecast has an MSE ratio of 1.000 with both DGP 1 and DGP 2, compared to the *opt. comb.: estimated* forecast's MSE ratio of 1.040 in DGP 1 and 1.038 in DGP 2.

Insert Table 3 here

However, making the DGP coefficient break smaller or larger than in the baseline case does lead to some changes in results – changes in line with the implications of the theory results in sections 2 and 3. With the smaller coefficient break (top panel of Table 3), using just the post-break sample to estimate the forecasting model yields a forecast much less (for smaller *P*) accurate than does using the full sample, with MSE ratios of roughly 1.17 for the $P = 1$ sample. The smaller break also gives the *opt. comb.: fixed R* forecast a larger advantage over the *rolling: fixed* R forecast. For instance, with DGP 1 and $P = 80$, the *opt. comb.: fixed R* forecast's MSE ratio is 1.009, compared to the *rolling: fixed* R forecast's MSE ratio of 1.045. One other change associated with making the DGP break smaller is that the BMA forecast with the fixed hyperparameter has a slight accuracy advantage over all the other feasible forecasts. Continuing with the DGP 1, $P = 80$ example, the *BMA, fixed prior variance* forecast has an MSE ratio of .990. Overall, with the smaller break, our proposed *opt. comb.: fixed R* forecast is nearly as accurate as the best-performing BMA forecast and as accurate as the next-best recursive forecast.

Making the DGP break larger also leads to some changes in results consistent with our theory findings. Broadly, the gains to combination over optimal sample determination, and the gains to optimal sample determination over using just a post-break window decline. For example, as shown for DGP 1 and $P = 1$ in the lower panel of Table 3, the MSE ratios of the *rolling: known R** and *opt. comb.: known* forecasts are .672 and .657, respectively, compared to the *rolling: known break R* forecasts MSE ratio of .675. Moreover, because the larger break is easier to empirically identify, the combination forecast based on the Andrews test-determined date is more accurate than the combination forecast based on the fixed break date of 40 observations prior to the forecast origin and the fixed rolling window forecast. For example, with DGP 2 and $P = 20$, the MSE ratios of the *rolling: fixed R*, *opt. comb.: estimated*, and *opt. comb.: fixed R* forecasts are, respectively, .679, .600, and .702. Finally, with the larger break, the BMA forecasts are sometimes a bit better and other times a bit worse than the *opt. comb.: estimated* forecast.[20] In the same example, the *BMA, fixed prior variance* forecast's MSE ratio is .577.

5. Application Results

To evaluate the empirical performance of the various forecast combination methods, we consider six different applications to U.S. data. In the first, we forecast quarterly U.S. GDP growth with one lag of growth, the spread between the 10-year Treasury bond yield and the 3-month Treasury bill rate, and the change in the 3-month rate. In the other five, we use common business cycle factors estimated as in Stock and Watson (2005) to forecast a selection of the monthly predictands considered by Stock and Watson: growth in payroll employment, growth in industrial production, the change in the unemployment rate, the change in the 3-month Treasury bill rate, and the change in CPI inflation. In each of these applications, the forecasting model includes six lags of the dependent variable and one lag of each of three common factors.[21]

---

[20] With the larger break, the best Bayesian forecast – unreported, but comparable to the *opt. comb.: estimated* forecast in the large break case – is one that picks a single break date, to maximize the marginal likelihood.

[21] The common factors are estimated with the principal component approach of Stock and Watson (2002, 2005), using a data set of 127 monthly series nearly identical to Stock and Watson's (2005). Following the specifications of Stock and Watson (2005), we first transformed the data for stationarity, screened for outliers, and standardized the data, and then computed principal components. We did so on a recursive basis, estimating different time series of factors at each forecast origin.

For all six applications, there is some evidence of historical instability in the relationship of interest.[22] For each application, a conventional Andrews (1993) test applied to the full sample of data rejects the null of stability (under both asymptotic and bootstrapped critical values).[23] For the CPI inflation application, the OLS-estimated break date is in 1974; in all other applications, the break date falls sometime in the early 1980s. Accordingly, our preceding theoretical and Monte Carlo results suggest that combining recursive and rolling forecasts may improve accuracy.

For these applications, we consider one-step ahead forecasts from 1985 through 2006:Q2 (GDP growth) or June 2006 (all other applications). In the GDP growth application, the model estimation sample begins with 1953:Q4; for the others, the estimation sample begins with July 1960.

The forecasts considered are the same as those included in the Monte Carlo analysis, with some minor modifications. The fixed rolling window forecasts use a window size of 10 years of data (40 observations for GDP growth, 120 observations for the other applications). In the break analysis, we impose a minimum break segment length of five years of data (20 observations for GDP growth, 60 observations for the other applications). We also, by necessity, drop consideration of the rolling forecasts based on the known post-break and known R* samples and combination based on the known optimal weight.

In line with common practice, we report our results in the form of MSEs relative to the MSE of a baseline forecast method, here taken to be the recursive forecast. For the recursive case, we report

[22] In addition, Estrella, et al. (2003) and Stock and Watson (2003), among others, report some evidence of instability in the relationship of GDP growth to interest rate term spreads.

[23] As first shown in Diebold and Chen (1996), Andrews (1993) tests applied to time series data tend to be over-sized, with the problem increasing in the degree of persistence in the data. Following Clark and McCracken (2006), in judging the significance of the break tests we consider critical values obtained with a wild bootstrap of a VAR in the series of interest.

the RMSE. For all others, we report the ratio of the MSE of the given forecast relative to the recursive forecast's MSE.

Insert Table 4 here

In broad terms, the application results in Table 4 are consistent with our theory and Monte Carlo results. In these applications, for which there is evidence of significant breaks, there is little or no advantage to using an optimal sample window over using just a post-break window. For example, in the GDP & interest rates application, the *rolling: post-break R* forecast's MSE ratio is .949, compared to the *rolling: estimated R** forecast's MSE ratio of .956. In most cases, the estimated optimal combination forecast improves on the accuracy of the optimal sample window forecast, but only modestly. In the same example, the *opt. comb.: estimated* forecast's MSE ratio is .951.

Reflecting the empirical difficulty of identifying breaks, using a fixed break date of 10 years prior to the forecast origin yields significantly more accurate forecasts. In the same example, the *opt. comb.: fixed R* forecast has an MSE ratio of .840. In the employment & factors application, the *opt. comb.: fixed R* forecast has an MSE ratio of .880, compared to .931, .963, and .941 for, respectively, the *rolling: post-break R*, *rolling: estimated R**, and *opt. comb.: estimated* forecasts. In these two applications, the coefficient break is apparently large enough that even the best-performing combination forecast is little or no more accurate than the *rolling: fixed R* forecast. However, in the other four applications, the *opt. comb.: fixed R* forecast improves upon the accuracy of the *rolling: fixed R* forecast. For example, in the 3-month T-bill & factors application, the *rolling: fixed R* and *opt. comb.: fixed R* forecasts have MSE ratios of, respectively, .988 and .926.

Finally, in these six applications, our proposed *opt. comb.: fixed R* forecast is generally, although not necessarily dramatically, more accurate than the BMA forecasts. In the GDP & interest rates application, the *BMA, fixed prior variance* forecast (in most of the applications, the fixed prior works better than the marginal likelihood-maximizing prior) has an MSE ratio of .958, compared to the *opt. comb.: fixed R* forecast's MSE ratio of .840. In the 3-month T-bill & factors application, the *BMA, fixed prior variance* and *opt. comb.: fixed R* forecasts have MSE ratios of, respectively, 1.009 and .926.

Overall, the results in Table 4 suggest that, in applications in which breaks may have occurred, combining forecasts from full sample and post-break sample model estimates can be a reasonably robust method for improving forecast accuracy. In light of the difficulty of empirically identifying breaks, unless the break evidence is overwhelming, it is likely better to impose an arbitrary break date such as 10 years prior to the forecast origin than to try to empirically identify the data. Such an approach appears to be at least competitive with alternatives such as Bayesian estimation and averaging of models with breaks.

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 theoretical results can be viewed as providing a frequentist justification for and approach to shrinkage; various Bayesian methods offer alternative, parallel justification. Our Monte Carlo results and results for a wide range of applications show that combining forecasts from models estimated with recursive and rolling samples consistently benefits forecast accuracy.

Clark: Economic Research Dept.; Federal Reserve Bank of Kansas City; 925 Grand; Kansas City, MO 64198; todd.e.clark@kc.frb.org.

McCracken (corresponding author): Board of Governors of the Federal Reserve System; 20th and Constitution N.W.; Mail Stop #61; Washington, D.C. 20551; michael.w.mccracken@frb.gov.

Appendix: Theoretical Results on the Bias-Variance Tradeoff

In this appendix we provide a Theorem that is used to derive the Corollaries in the text. Along with proofs of the Corollaries, as an aid in the proofs, an intermediate appendix Corollary is provided. 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. For brevity, in the statement of Theorem 1, we define the function $g(t/T) = 1(t/T \geq T_B/T)\Delta\beta$ and its weak limit $g(s)$. For any $(m \times n)$ matrix A with elements $a_{i,j}$ and column vectors a_j, let $vec(A)$ denote the $(mn \times 1)$ vector $[a'_1, a'_2, ..., a'_n]'$, let $|A|$ denote the max norm and define $\sup_t = \sup_{T\leq t\leq T+P}$. For the recursive scheme let $H_{T,R}(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{T,s}u_{T,s+\tau}) = (t^{-1}\sum_{s=1}^{t-\tau} h_{T,s+\tau})$, $B_{T,R}(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{T,s}x'_{T,s})^{-1}$ and $D_{T,R}(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{T,s}x'_{T,s}g(s/T))$. For the rolling scheme let $H_{T,L}(t) = (R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau} x_{T,s}u_{T,s+\tau}) = (R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau} h_{T,s+\tau})$, $B_{T,L}(t) = (R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau} x_{T,s}x'_{T,s})^{-1}$ and $D_{T,L}(t) = (R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau} x_{T,s}x'_{T,s}g(s/T))$. In this notation we obtain $\hat{\beta}_{R,t} - \beta^* = T^{-1/2}B_{T,R}(t)D_{T,R}(t) + B_{T,R}(t)H_{T,R}(t)$ and $\hat{\beta}_{L,t} - \beta^* = T^{-1/2}B_{T,L}(t)D_{T,L}(t) + B_{T,L}(t)H_{T,L}(t)$.

<u>Assumption 1:</u> The DGP satisfies $y_{T,t+\tau} = x'_{T,t}\beta^*_{T,t} + u_{T,t+\tau}$, $\beta^*_{T,t} = \beta^* + T^{-1/2}1(t \geq T_B)\Delta\beta$ for all $t = 1,...,T,...T+P$, such that $\lim_{T\to\infty} T_B/T = \lambda_B > 0$.

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

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

Assumption 4: For $s \in (0, 1 + \lambda_P]$, (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, $\sum_{t=T}^{T+P} (\hat{u}^2_{R,t+\tau} - \hat{u}^2_{W,t+\tau}) \to_d$

$$\{-2\int_1^{1+\lambda_P} (1 - \alpha(s))[s^{-1}W(s) - \lambda_R^{-1}(s)(W(s) - W(s - \lambda_R(s)))]' V^{1/2} B V^{1/2} dW(s)$$

$$+ \int_1^{1+\lambda_P} (1 - \alpha^2(s)) s^{-2} W(s)' V^{1/2} B V^{1/2} W(s) ds$$

$$- \int_1^{1+\lambda_P} (1 - \alpha(s))^2 \lambda_R^{-2} (W(s) - W(s - \lambda_R(s)))' V^{1/2} B V^{1/2} (W(s) - W(s - \lambda_R(s)))] ds \}$$

$$- 2\int_1^{1+\lambda_P} \alpha(s)(1 - \alpha(s)) s^{-1} \lambda_R^{-1}(s) W(s)' V^{1/2} B V^{1/2} (W(s) - W(s - \lambda_R(s))) ds \}$$

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

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

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

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

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

$$- \int_1^{1+\lambda_P} (1 - \alpha(s))g(s)' V^{1/2}[s^{-1}W(s) - \lambda_R^{-1}(s)(W(s) - W(s - \lambda_R(s)))]ds \}$$

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

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

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

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

Proof of Theorem 1: Straightforward but tedious algebra along the lines of West (1996) and Clark and McCracken (2001) reveals that

(6) $\sum_{t=T}^{T+P} (\hat{u}_{R,t+\tau}^2 - \hat{u}_{W,t+\tau}^2) =$

$$\{-2\sum_{t=T}^{T+P} (1 - \alpha_t)(T^{-1/2}h_{T,t+\tau}')B(T^{1/2}[H_{T,R}(t) - H_{T,L}(t)])$$

$$- 2T^{-1}\sum_{t=T}^{T+P} \alpha_t(1 - \alpha_t)H_{T,R}'(t)BH_{T,L}(t))$$

$$+ T^{-1}\sum_{t=T}^{T+P} [(1 - \alpha_t^2)(T^{1/2}H_R'(t))B(T^{1/2}H_{T,R}(t)) - (1 - \alpha_t)^2(T^{1/2}H_{T,L}'(t))B(T^{1/2}H_{T,L}(t))]\}$$

$$+ 2\{-\sum_{t=T}^{T+P} (1 - \alpha_t)(T^{-1/2}h_{T,t+\tau}')[(t^{-1}\sum_{s=1}^{t-\tau} g(s/T)) - (R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau} g(s/T))]$$

$$+ T^{-1}\sum_{t=T}^{T+P} [(1 - \alpha_t^2)(T^{1/2}H_{T,R}'(t))(t^{-1}\sum_{s=1}^{t-\tau} g(s/T))$$

$$- (1 - \alpha_t)^2(T^{1/2}H_{T,L}'(t))(R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau} g(s/T))]$$

$$- T^{-1}\sum_{t=T}^{T+P} \alpha_t(1 - \alpha_t)[(T^{1/2}H_{T,R}'(t))(R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau} g(s/T))$$

$$- (T^{1/2}H_{T,L}'(t))(t^{-1}\sum_{s=1}^{t-\tau} g(s/T))]$$

$$- T^{-1}\sum_{t=T}^{T+P} (1 - \alpha_t)g(t/T)'[(T^{1/2}H_{T,R}(t)) - (T^{1/2}H_{T,L}(t))]\}$$

$$+ \{ -2T^{-1}\sum_{t=T}^{T+P}(1-\alpha_t)g(t/T)'B^{-1}[(t^{-1}\sum_{s=1}^{t-\tau}g(s/T)) - (R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}g(s/T))]$$

$$+ T^{-1}\sum_{t=T}^{T+P}[(1-\alpha_t^2)(t^{-1}\sum_{s=1}^{t-\tau}g(s/T))'B^{-1}(t^{-1}\sum_{s=1}^{t-\tau}g(s/T))$$

$$-(1-\alpha_t)^2(R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}g(s/T))'B(R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}g(s/T))]$$

$$-2T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(t^{-1}\sum_{s=1}^{t-\tau}g(s/T))'B^{-1}(R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}g(s/T))\} + o_p(1).$$

For the remainder we show that the expansion in (6) converges in distribution to the term provided in the Theorem. To do so recall that Assumption 4 maintains $R_t/T \Rightarrow \lambda_R(s)$ and $\alpha_t \Rightarrow \alpha(s)$. With that in mind, continuity and Assumption 1 imply $g(t/T) \Rightarrow g(s)$, $t^{-1}\sum_{s=1}^{t-\tau}g(s/T) \Rightarrow s^{-1}(\int_0^s g(r)dr)$ and $R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}g(s/T) \Rightarrow \lambda_R^{-1}(s)(\int_{s-\lambda_R(s)}^s g(r)dr$. Assumptions 3 (a) - (d) then imply both $T^{1/2}H_{T,R}(t) \Rightarrow s^{-1}V^{1/2}W(s)$ and $T^{1/2}H_{T,L}(t) \Rightarrow \lambda_R^{-1}(s)V^{1/2}(W(s) - W(s-\lambda_R(s)))$. The continuous mapping theorem then provides the desired results for the second and third contributions to the first bracketed term, for the second, third and fourth contributions to the second bracketed term and all three contributions to the third.

The remaining two contributions (the first in each of the first two bracketed terms), are each weighted sums of increments $h_{T,t+\tau}$. Consider the first contribution to the second bracketed term. Since this increment satisfies Assumption 3 (d) and has an associated long-run variance V, we can apply Theorem 4.1 of de Jong and Davidson (2000) directly to obtain the desired convergence in distribution

$$-\sum_{t=T}^{T+P}(1-\alpha_t)(T^{-1/2}h_{T,t+\tau}')[(t^{-1}\sum_{s=1}^{t-\tau}g(s/T)) - (R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}g(s/T))]$$

$$\to_d -\int_1^{1+\lambda_P}(1-\alpha(s))[s^{-1}(\int_0^s g(r)dr) - \lambda_R^{-1}(s)(\int_{s-\lambda_R(s)}^s g(r)dr)]'V^{1/2}dW(s).$$

For the first contribution to the first bracketed term additional care is needed. Again, since the increments satisfy Assumption 3 (d) with long-run variance V we can apply Theorem 4.1 of de Jong and Davidson (2000) to obtain

$$-2\sum_{t=T}^{T+P}(1 - \alpha_t)(T^{-1/2}h'_{T,t+\tau})B(T^{1/2}[H_{T,R}(t) - H_{T,L}(t)])$$

$$\to_d -2\int_1^{1+\lambda_P}(1 - \alpha(s))[s^{-1}W(s) - \lambda_R^{-1}(s)(W(s) - W(s - \lambda_R(s)))]'V^{1/2}BV^{1/2}dW(s) + \Lambda .$$

Note the addition of the drift term Λ. To obtain the desired result we must show that this term is zero. A detailed proof is provided in Lemma A6 of Clark and McCracken (2005) – albeit under the technical conditions provided in Hansen (1992) rather than those provided here. Rather than repeat the proof we provide an intuitive argument. Note that $H_{T,R}(t) = t^{-1}\sum_{s=1}^{t-\tau}h_{T,s+\tau}$ while $H_{T,L}(t) = R_t^{-1}\sum_{s=t-\tau-R_t+1}^{t-\tau}h_{T,s+\tau}$. In particular note the range of summation. Since Assumption 3 (b) maintains that the increments of the stochastic integral $h_{T,t+\tau}$ form an MA($\tau - 1$) we find that $h_{T,t+\tau}$ is uncorrelated with every element of $H_{T,R}(t)$ and $H_{T,L}(t)$. Since Λ captures the contribution to the mean of the limiting distribution due to covariances between the increments $h_{T,t+\tau}$ and the elements of $H_{T,R}(t) - H_{T,L}(t)$ we know that $\Lambda = 0$ and the proof is complete.

Proof of Corollary 2.1: Follows as a special case of Corollary 3.1 when $\alpha_t = 0$ for all t.

Proof of Corollary 3.1: For both cases (a) and (b), note that the expectation of the second bracketed term $\{.\}$, from Theorem 1, is zero. Note also that the first bracketed term does not depend upon λ_B and hence its expectation is the same for both cases. Taking the expectation of this term we obtain

$$\{ -2\int_1^{1+\lambda_P} (1 - \alpha(s)) tr(V^{1/2}BV^{1/2}E([s^{-1}W(s) - \lambda_R^{-1}(s)(W(s) - W(s - \lambda_R(s)))]dW'(s)))$$

$$+ \int_1^{1+\lambda_P} (1 - \alpha^2(s)) s^{-2} tr(V^{1/2}BV^{1/2}E(W(s)W(s)'))ds$$

$$- \int_1^{1+\lambda_P} (1 - \alpha(s))^2 \lambda_R^{-2}(s) tr(V^{1/2}BV^{1/2}E((W(s) - W(s - \lambda_R(s))) \times$$

$$(W(s) - W(s - \lambda_R(s)))')ds$$

$$- 2\int_1^{1+\lambda_P} \alpha(s)(1 - \alpha(s)) s^{-1} \lambda_R^{-1}(s) tr(V^{1/2}BV^{1/2}E((W(s) - W(s - \lambda_R(s)))W(s)'))ds \}$$

$$= 0 + tr(BV) \{ \int_1^{1+\lambda_P} (1 - \alpha^2(s)) s^{-1} ds - \int_1^{1+\lambda_P} (1 - \alpha(s))^2 \lambda_R^{-2}(s)(s - \lambda_R(s))ds$$

$$- 2\int_1^{1+\lambda_P} \alpha(s)(1 - \alpha(s)) s^{-1} \lambda_R^{-1}(s)(s - \lambda_R(s))ds \}$$

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

(a) We now derive the expectation of the third bracketed term under the assumption of a single break of magnitude $\Delta\beta$ at time λ_B given $\lambda_R(s) > s - \lambda_B$ all $s \in [1, 1 + \lambda_P]$. Under this restriction first note that

$$g(s) = \begin{cases} \Delta\beta & s > \lambda_B \\ 0 & s \leq \lambda_B \end{cases}, \qquad \int_0^s g(r)dr = \int_{s-\lambda_R(s)}^s g(r)dr = \begin{cases} (s - \lambda_B)\Delta\beta & s > \lambda_B \\ 0 & s \leq \lambda_B \end{cases}.$$

After taking expectations, direct substitution and algebra then provides

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

$$+ \int_1^{1+\lambda_P}(s - \lambda_B)^2((1 - \alpha^2(s))s^{-2} - (1 - \alpha(s))^2\lambda_R^{-2}(s))ds$$

$$- 2\int_1^{1+\lambda_P}\alpha(s)(1 - \alpha(s))s^{-1}\lambda_R^{-1}(s)(s - \lambda_B)^2 ds\,\}$$

$$= \Delta\beta' B^{-1}\Delta\beta\int_1^{1+\lambda_P}(1 - \alpha(s))(s - \lambda_R(s))(s - \lambda_B)\times$$

$$\left(\frac{(s - \lambda_B)(\alpha(s)(s - \lambda_R(s)) - (s + \lambda_R(s))) + 2s\lambda_R(s))}{s^2\lambda_R^2(s)}\right)ds\,.$$

(b) We now derive the expectation of the third bracketed term under the assumption of a single break of magnitude $\Delta\beta$ at time λ_B given $\lambda_R(s) \leq s - \lambda_B$ all $s \in [1, 1 + \lambda_P]$. Under this restriction first note that

$$g(s) = \begin{cases}\Delta\beta & s > \lambda_B \\ 0 & s \leq \lambda_B\end{cases}, \quad \int_0^s g(r)dr = \begin{cases}(s - \lambda_B)\Delta\beta & s > \lambda_B \\ 0 & s \leq \lambda_B\end{cases},$$

$$\int_{s-\lambda_R(s)}^{s} g(r)dr = \begin{cases}\lambda_R(s)\Delta\beta & s > \lambda_B \\ 0 & s \leq \lambda_B\end{cases}.$$

After taking expectations, direct substitution and algebra then provides

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

$$- 2\int_1^{1+\lambda_P}\alpha(s)(1 - \alpha(s))s^{-1}(s - \lambda_B)ds\,\}$$

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

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

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

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

Proof of Appendix Corollary: We will first derive the reaction function $\lambda_R^*(s,\alpha(s))$. We do so by maximizing the sum of the arguments of the integrals in Corollary 3.1, for each fixed s, keeping in mind the piecewise nature of the objective function at $\lambda_R = s - \lambda_B$. That $\lambda_R^*(s,\alpha(s))$ is invariant when $\alpha = 1$ arises since no weight is being placed on the rolling component of the combined forecast. That $\lambda_R^*(s,\alpha(s))$ will never be less than $s - \lambda_B$ arises since the bias term in the expansion of Corollary 3.1 (b) does not depend upon $\lambda_R^*(s,\alpha(s))$ and the variance term in the same expansion is monotone increasing in $\lambda_R^*(s,\alpha(s))$.

For the other two components of the reaction function we need only consider optimizing the expansion of Corollary 3.1 (a). Hence the derivation is based upon maximizing

$$(7)\ \ tr(BV)(1 - \alpha)^2(\frac{1}{s} - \frac{1}{\lambda_R})$$

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

Note that for brevity, since the optimization is performed holding the index s fixed, we omit it from α and λ_R. Differentiating (7) with respect to λ_R we obtain

$$\text{FOC } \lambda_R : \frac{tr(BV)(1 - \alpha)^2}{\lambda_R^2} + \Delta\beta' B^{-1} \Delta\beta \{ -\frac{2(1 - \alpha)(s - \lambda_B)}{\lambda_R^2}$$

$$+ \frac{2(s - \lambda_B)^2(1 - \alpha)^2}{\lambda_R^3} + \frac{2\alpha(1 - \alpha)(s - \lambda_B)^2}{s\lambda_R^2} \}$$

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

$$+ 2s(s - \lambda_B)^2(1 - \alpha) + 2\alpha\lambda_R(s - \lambda_B)^2 \}]$$

$$\text{SOC } \lambda_R : -\frac{2tr(BV)(1 - \alpha)^2}{\lambda_R^3} + \Delta\beta' B^{-1} \Delta\beta \{ \frac{4(1 - \alpha)(s - \lambda_B)}{\lambda_R^3}$$

$$- \frac{6(s - \lambda_B)^2(1 - \alpha)^2}{\lambda_R^4} - \frac{4\alpha(1 - \alpha)(s - \lambda_B)^2}{s\lambda_R^3} \}$$

$$= -2(\frac{1 - \alpha}{s\lambda_R^4}) [tr(BV) s\lambda_R (1 - \alpha) - \Delta\beta' B^{-1} \Delta\beta \{ 2s\lambda_R(s - \lambda_B)$$

$$- 3s(s - \lambda_B)^2(1 - \alpha) - 2\alpha\lambda_R(s - \lambda_B)^2 \}]$$

To solve for the (potential) interior solution we set the FOC equal to zero and solve for λ_R.

$$0 = tr(BV)s\lambda_R(1 - \alpha) + \Delta\beta' B^{-1}\Delta\beta \{-2s\lambda_R(s - \lambda_B) + 2s(s - \lambda_B)^2(1 - \alpha) + 2\alpha\lambda_R(s - \lambda_B)^2\}$$

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

or

$$\lambda_R^*(\alpha) = \frac{2s(1 - \alpha(s))(s - \lambda_B)^2 \frac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)}}{2(s - \lambda_B)(s - \alpha(s)(s - \lambda_B))\frac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)} - s(1 - \alpha(s))}.$$

Straightforward algebra reveals both the constraints ensure that $s - \lambda_B \leq \lambda_R^*(\alpha) \leq s$ and that the SOC is negative when evaluated at $\lambda_R^*(\alpha)$.

Now consider the reaction function $\alpha^*(s, \lambda_R(s))$. When $s = \lambda_R$ we find that the optimal value of α is invariant because the recursive and rolling forecasts are identical. For $s < \lambda_R$, in contrast to the previous derivation, we must explicitly consider interior solutions for the value of α that maximizes not only equation (7) but also its equivalent from Corollary 3.1 (b)

$$(8) \quad tr(BV)(1 - \alpha)^2(\frac{1}{s} - \frac{1}{\lambda_R}) + \Delta\beta' B^{-1}\Delta\beta(1 - \alpha^2)(\frac{\lambda_B^2}{s^2}).$$

Differentiating (7) and (8) with respect to α we obtain

$$\text{FOC (7) } \alpha: -2tr(BV)(1 - \alpha)(\frac{1}{s} - \frac{1}{\lambda_R}) + \Delta\beta' B^{-1}\Delta\beta \{-\frac{2(s - \lambda_B)(s - \lambda_R)}{s\lambda_R}$$

$$+ \ 2(s - \lambda_B)^2(-\frac{\alpha}{s^2} + \frac{(1 - \alpha)}{\lambda_R^2}) - \frac{2(1 - 2\alpha)(s - \lambda_B)^2}{s\lambda_R} \}$$

$$= (\frac{2}{s^2\lambda_R^2})[\, tr(BV)(1 - \alpha)s\lambda_R(s - \lambda_R) + \Delta\beta' B^{-1}\Delta\beta \{-s\lambda_R(s - \lambda_B)(s - \lambda_R)$$

$$+ \ (s - \lambda_B)^2((1 - \alpha)s^2 - \alpha\lambda_R^2) - (1 - 2\alpha)s\lambda_R(s - \lambda_B)^2 \}$$

SOC (7) $\alpha:\ 2tr(BV)(\frac{1}{s} - \frac{1}{\lambda_R}) + \Delta\beta' B^{-1}\Delta\beta \{-2(s - \lambda_B)^2(\frac{1}{s^2} + \frac{1}{\lambda_R^2}) + \frac{4(s - \lambda_B)^2}{s\lambda_R} \}$

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

FOC (8) $\alpha:\ -2tr(BV)(1 - \alpha)(\frac{1}{s} - \frac{1}{\lambda_R}) - \{\frac{2\alpha\lambda_B^2\Delta\beta' B^{-1}\Delta\beta}{s^2}\}$

SOC (8) $\alpha:\ 2tr(BV)(\frac{1}{s} - \frac{1}{\lambda_R}) - \{\frac{2\lambda_B^2\Delta\beta' B^{-1}\Delta\beta}{s^2}\}$

To solve for the (potential) interior solutions we set each FOC equal to zero and solve for α. Doing so for FOC (7) we find that

$$0 = tr(BV)(1 - \alpha)s\lambda_R(s - \lambda_R) + \Delta\beta' B^{-1}\Delta\beta \{-s\lambda_R(s - \lambda_B)(s - \lambda_R)$$

$$+ \ (s - \lambda_B)^2((1 - \alpha)s^2 - \alpha\lambda_R^2) - (1 - 2\alpha)s\lambda_R(s - \lambda_B)^2$$

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

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

or

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

Using similar arguments for FOC (8) we find that $\alpha^*(\lambda_R) = \dfrac{s(s - \lambda_R)}{s(s - \lambda_R) + (\frac{\Delta\beta' B^{-1}\Delta\beta}{tr(BV)})\lambda_B^2\lambda_R}$.

Straightforward algebra reveals that both second order conditions are uniformly negative over their respective ranges of α and hence the proof is complete.

Proof of Corollary 2.2: Follows as a special case of the Appendix Corollary when $\alpha_t = 0$ for all t.

Proof of Corollary 3.2: The result follows from combining the two reaction functions from the Appendix Corollary.

References

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

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

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

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

Clark, T.E. and M.W. McCracken, "Evaluating Direct Multi-Step Forecasts", Econometric Reviews, 24 (2005), 369-404.

Clark, T.E. and M.W. McCracken, "Averaging Forecasts from VARs with Uncertain Instabilities", Journal of Applied Econometrics, forthcoming, 2006.

Cogley, T. and T.J. Sargent, "Drifts and Volatilities Monetary Policies and Outcomes in the Post World War II U.S.," Review of Economic Dynamics, 8 (2005), 262-302.

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

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters, "On the Fit and Forecasting Performance of New Keynesian Models," Journal of Business and Economic Statistics, 25 (2007), 123-162.

Diebold, F.X. and C. Chen, "Testing Structural Stability with Endogenous Breakpoint A Size Comparison of Analytic and Bootstrap Procedures," Journal of Econometrics, 70 (1996), 221-241.

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

Elliott, G., "Forecasting When There is a Single Break," manuscript, UCSD, 2005.

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

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

Fama E., and J. MacBeth, "Risk, Return, and Equilibrium Empirical Tests," Journal of Political Economy, 81 (1973), 607-636.

Hansen, B.E., “Convergence to Stochastic Integrals for Dependent Heterogeneous Processes,” Econometric Theory, 8 (1992), 489-500.

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

Hansen, B.E., “Testing for Structural Change in Conditional Models,” Journal of Econometrics, 97 (2000), 93-115.

Hooker, M.A., “Are Oil Shocks Inflationary? Asymmetric and Nonlinear Specifications versus Changes in Regime,” Journal of Money, Credit, and Banking, 34 (2002), 540-61.

Hultblad, B. and S. Karlsson, “Bayesian Simultaneous Determination of Structural Breaks and Lag Lengths,” manuscript, Stockholm School of Economics, 2006.

Inoue, A. and L. Kilian, “On the Selection of Forecasting Models,” ECB Working Paper 214, 2003.

Inoue, A. and B. Rossi, “Recursive Predictability Tests for Real-Time Data,” Journal of Business and Economic Statistics, 23 (2005), 336-45.

Koop, G. and S. Potter, “Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging,” Econometrics Journal, 7 (2004), 550-65.

Maheu, J.M. and S. Gordon, “Learning, Forecasting and Structural Breaks”, Journal of Applied Econometrics, forthcoming, 2007.

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

Paye, B.S. and A. Timmermann, “Instability of Return Prediction Models,” Journal of Empirical Finance, 13 (2006), 274-315.

Pesaran, M.H. and A. Timmermann, “Market Timing and Return Prediction Under Model Instability,” Journal of Empirical Finance, 9 (2002), 495-510.

Pesaran, M.H. and A. Timmermann, “Selection of Estimation Window in the Presence of Breaks,” Journal of Econometrics, 137 (2007), 134-161.

Pesaran, M.H., D. Pettenuzzo and A. Timmermann, “Forecasting Time Series Subject to Structural Breaks,” Review of Economic Studies, 73 (2006), 1057-1084.

Rapach, D.E., and M.E. Wohar, “Structural Breaks and Predictive Regression Models of Aggregate U.S. Stock Returns,” Journal of Financial Econometrics, 4 (2006), 238-274.

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

Stock, J.H., and M.W. Watson, "Macroeconomic Forecasting Using Diffusion Indexes," Journal of Business and Economic Statistics, 20 (2002), 147-162.

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

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

Stock, J.H., and M.W. Watson, "An Empirical Comparison of Methods for Forecasting Using Many Predictors," manuscript, Harvard University, 2005.

Swanson, N.R., "Money and Output Viewed Through A Rolling Window," Journal of Monetary Economics, 41 (1998), 455-474.

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

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

Yao, Y-C., “Estimating the Number of Change-Points Via Schwarz’ Criterion,” Statistics and Probability Letters, 6 (1988), 181-89.

Table 1: Summary of Forecast Approaches

approach	explanation
recursive	coefficient estimates based on all available data
rolling: known break R	coefficient estimates based on post-break sample, using the known break date
rolling: known R*	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
opt. comb: 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
rolling: fixed R	coefficient estimates based on R most recent observations, with R = 40
rolling: post-break R	coefficient estimates based on post-break sample, using sup Wald-based estimates of the break point and sample moment estimates
rolling: estimated R*	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.
opt. comb.: 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 (2.5% sig.level) and the estimated date of the break
opt. comb.: fixed R	combination of the recursive forecast and a forecast based on rolling parameter estimates from the R most recent observations, with R = 40, and weights estimated using (5)
BMA, fixed prior variance	Bayesian model average of forecasts from models allowing a single break at an unknown date, within a range of observations 21 and t-20. The prior probability on each model or forecast is 1/number of possible break dates. For each model, the prior on the pre-break coefficients is loose, while the prior on the change in coefficients at the break date is informative, with a mean of zero.
BMA, optimized prior variance	same as above, except that the hyperparameter determining the informativeness of the prior on the break size is data-determined, to maximize the marginal likelihood of the average forecast.

Table 2: Baseline Monte Carlo Results, Average MSEs
(average MSE for recursive, and ratio of average MSE to recursive average for other forecasts)

	Break point: observation 80							
	DGP 1				DGP 2			
	$P=1$	$P=20$	$P=40$	$P=80$	$P=1$	$P=20$	$P=40$	$P=80$
recursive	13.796	13.321	12.954	12.479	2.199	2.102	2.043	1.961
rolling: known break R	.979	.925	.916	.914	.976	.929	.914	.912
rolling: known R*	.950	.917	.911	.912	.939	.922	.914	.914
opt. comb.: known	.900	.884	.888	.898	.882	.877	.880	.892
rolling: fixed R	.940	.916	.921	.945	.949	.919	.921	.946
rolling: post-break R	1.007	.986	.976	.961	.994	.988	.972	.956
rolling: estimated R*	.992	.964	.956	.947	.971	.961	.950	.941
opt. comb.: estimated	.980	.962	.955	.946	.960	.951	.941	.934
opt. comb.: fixed R	.940	.914	.913	.930	.944	.913	.908	.925
BMA, fixed prior variance	.930	.916	.915	.917	.882	.875	.880	.892
BMA, optimized prior variance	.964	.940	.935	.931	.914	.896	.893	.898
	Break point: observation 60							
	DGP 1				**DGP 2**			
	$P=1$	$P=20$	$P=40$	$P=80$	$P=1$	$P=20$	$P=40$	$P=80$
recursive	12.484	12.256	12.111	11.827	2.024	1.942	1.900	1.844
rolling: known break R	.936	.929	.930	.935	.909	.920	.923	.931
rolling: known R*	.933	.928	.930	.936	.916	.925	.928	.936
opt. comb.: known	.915	.915	.920	.929	.893	.906	.913	.925
rolling: fixed R	.936	.952	.965	.989	.909	.944	.965	.992
rolling: post-break R	1.001	.987	.979	.974	.966	.974	.971	.968
rolling: estimated R*	.980	.970	.965	.964	.949	.956	.956	.957
opt. comb.: estimated	.978	.969	.964	.963	.940	.949	.950	.953
opt. comb.: fixed R	.924	.935	.946	.965	.895	.923	.939	.958
BMA, fixed prior variance	.932	.933	.934	.940	.894	.906	.914	.926
BMA, optimized prior variance	.954	.948	.947	.949	.904	.911	.917	.928

Notes:
1. DGPs 1 and 2 are defined in Section 4.1. The forecast approaches are defined in Table 1.
2. The total number of observations in each experiment is 180. Forecasting begins with observation 101. Results are reported for forecasts evaluated from period 101 through 180. The break in the DGP occurs at observation 80 (i.e., $\lambda_B = .8$) in the experiment results reported in the upper panel and observation 60 ($\lambda_B = .6$) in the experiment results reported in the lower panel.
3. The table entries are based on averages of forecast MSEs across 5000 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 Smaller and Larger Breaks, Average MSEs
(average MSE for recursive, and ratio of average MSE to recursive average for other forecasts)

	Smaller break at observation 80							
	DGP 1				DGP 2			
	$P=1$	$P=20$	$P=40$	$P=80$	$P=1$	$P=20$	$P=40$	$P=80$
recursive	11.607	11.456	11.351	11.183	1.835	1.790	1.767	1.738
rolling: known break R	1.166	1.077	1.046	1.021	1.177	1.092	1.057	1.029
rolling: known R*	1.003	1.006	1.002	.996	1.000	1.000	1.000	.998
opt. comb.: known	.989	.982	.982	.981	.984	.982	.981	.983
rolling: fixed R	1.034	1.029	1.033	1.045	1.044	1.037	1.043	1.057
rolling: post-break R	1.069	1.064	1.061	1.054	1.072	1.074	1.070	1.061
rolling: estimated R*	1.043	1.040	1.038	1.032	1.040	1.040	1.038	1.033
opt. comb.: estimated	1.044	1.040	1.038	1.033	1.037	1.038	1.035	1.031
opt. comb.: fixed R	1.007	1.000	1.002	1.009	1.006	1.000	1.002	1.009
BMA, fixed prior variance	.993	.990	.990	.990	.997	.995	.996	.997
BMA, optimized prior variance	1.004	1.001	1.000	.999	.993	.991	.990	.989
	Larger break at observation 80							
	DGP 1				**DGP 2**			
	$P=1$	$P=20$	$P=40$	$P=80$	$P=1$	$P=20$	$P=40$	$P=80$
recursive	19.942	18.422	17.291	15.973	3.608	3.323	3.128	2.847
rolling: known break R	.675	.667	.685	.714	.595	.590	.599	.629
rolling: known R*	.672	.669	.687	.715	.610	.601	.608	.635
opt. comb.: known	.657	.658	.678	.710	.574	.578	.591	.625
rolling: fixed R	.788	.733	.728	.759	.778	.679	.653	.680
rolling: post-break R	.723	.701	.709	.728	.615	.611	.614	.638
rolling: estimated R*	.725	.701	.708	.728	.634	.618	.619	.643
opt. comb.: estimated	.711	.692	.702	.724	.604	.600	.606	.634
opt. comb.: fixed R	.815	.748	.735	.760	.812	.702	.664	.683
BMA, fixed prior variance	.834	.803	.796	.793	.582	.577	.590	.623
BMA, optimized prior variance	.772	.723	.724	.738	.607	.592	.600	.629

Notes:
1. DGPs 1 and 2 are defined in Section 4.1. In the experiments in the upper panel, the breaks imposed in the DGPs are 1/2 the size of those imposed in the baseline experiments. In the experiments in the lower panel, the breaks imposed in the DGPs are twice the size of those imposed in the baseline experiments. The forecast approaches are defined in Table 1.
2. The total number of observations in each experiment is 180. Forecasting begins with observation 101. Results are reported for forecasts evaluated from period 101 through 180. The break in the DGP occurs at observation 80 (i.e., $\lambda_B = .8$).
3. The table entries are based on averages of forecast MSEs across 5000 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: Application Results, 1985-2006 Forecast Accuracy

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

	GDP & interest rates	**Employment & factors**
recursive	2.384	1.225
rolling: fixed R	.831	.894
rolling: post-break R	.949	.931
rolling: estimated R*	.956	.963
opt. comb.: estimated	.951	.941
opt. comb.: fixed R	.840	.880
BMA, fixed prior variance	.958	.948
BMA, optimized prior variance	.992	.986
	Ind. prod. & factors	**Unemployment rate & factors**
recursive	6.141	.133
rolling: fixed R	.997	.969
rolling: post-break R	.998	1.003
rolling: estimated R*	1.011	1.003
opt. comb.: estimated	.992	.981
opt. comb.: fixed R	.974	.961
BMA, fixed prior variance	.975	.990
BMA, optimized prior variance	.999	.988
	3-month T-bill & factors	**CPI inflation & factors**
recursive	.207	2.522
rolling: fixed R	.988	.995
rolling: post-break R	1.190	1.067
rolling: estimated R*	1.103	1.035
opt. comb.: estimated	1.074	1.039
opt. comb.: fixed R	.926	.964
BMA, fixed prior variance	1.009	1.015
BMA, optimized prior variance	1.064	.992

Notes:
1. Details of the six applications (data, forecast model specification, etc.) are provided in Section 5. In all cases, the units of the predictand are annualized percentage points.
2. The forecast approaches listed in the first column are defined in Table 1. Note that, for the fixed R rolling forecasts, R = 40 for the (quarterly) GDP application and R = 120 for the other (monthly) applications. For the forecasts based on break date estimates, the minimum sample window allowed is 20 observations in the (quarterly) GDP application and 60 observations in the other (monthly) applications.