Optimal Forecasts in the Presence of Structural Breaks*

M. Hashem Pesaran
University of Southern California
and Trinity College, Cambridge

Andreas Pick
Erasmus University Rotterdam
and De Nederlandsche Bank

Mikhail Pranovich
Joint Vienna Institute

February 9, 2013

Abstract

This paper considers the problem of forecasting under continuous and discrete structural breaks and proposes weighting observations to obtain optimal forecasts in the MSFE sense. We derive optimal weights for one step ahead forecasts. Under continuous breaks, our approach largely recovers exponential smoothing weights. Under discrete breaks, we provide analytical expressions for optimal weights in models with a single regressor, and asymptotically valid weights for models with more than one regressor. It is shown that in these cases the optimal weight is the same across observations within a given regime and differs only across regimes. In practice, where information on structural breaks is uncertain, a forecasting procedure based on robust optimal weights is proposed. The relative performance of our proposed approach is investigated using Monte Carlo experiments and an empirical application to forecasting real GDP using the yield curve across nine industrial economies.

Keywords: Forecasting, structural breaks, optimal weights, robust optimal weights, exponential smoothing
JEL codes: C22, C53

*We would like to thank Don Harding and Ana Galvao for interesting discussions and two anonymous referees for their comments. Additionally, we are grateful for helpful comments from participants at the 31st Annual International Symposium on Forecasting in Prague, the 7th ECB workshop on forecasting techniques, the MMF/CEF workshop at Brunel University, the NESG meeting in Groningen, ESEM in Malaga, and a seminar at ESSEC.

1 Introduction

It is now widely recognized that parameter instability is an important source of forecast failure in macroeconomics and finance as documented by Pesaran and Timmermann (2002), Pesaran, Pettenuzzo, Timmermann (2006), Koop and Potter (2007), Giacomini and Rossi (2009), Inoue and Rossi (2011), among others. Clements and Hendry (1999, 2006) and Rossi (2011) provide reviews. Broadly speaking, there are two basic approaches to modeling parameter instability: parameters are assumed to change either at discrete time intervals or continuously. Under the former, break dates are estimated and forecasts are typically constructed using the post-break observations.[1] Assuming that the break dates are accurately estimated, the forecasts based on observations after the last break are likely to be unbiased. However, as pointed out by Pesaran and Timmermann (2007), forecasts from the post-break window may not minimize the mean square forecast error (MSFE) as the estimation uncertainty may be large due to the relatively short post-break window. For this reason Pesaran and Timmermann (2007) suggest an optimal estimation window that may include pre-break observations. When the time and size of the break is uncertain, Pesaran and Timmermann (2007) consider averaging forecasts across estimation windows (AveW), which, as Pesaran and Pick (2011) show, improves forecasts without relying on estimates of break dates and sizes.

Under the continuously changing parameter model, the breaks are assumed to occur every period, and observations are down-weighted to take account of the slowly changing nature of the parameters. Within this framework, a prominent approach is exponential smoothing (ExpS), first proposed by Holt (1957) and Brown (1959). Other approaches using Kalman filters have also been proposed as generalizations of ExpS. Hyndman, Koehler, Ord, and Snyder (2008) provide a comprehensive survey.

In this paper, we develop a unified approach to obtaining optimal forecasts under both types of structural breaks, focusing on one-step-ahead forecasts. We consider forecasts based on weighted observations as in the ExpS approach but derive weights that are optimal in the sense that the resulting forecasts minimize the MSFE. In the case of continuous breaks, the optimal weights approximate ExpS weights if T is large and the downweighting parameter of ExpS not too close to unity. In contrast, when the breaks are assumed to occur at discrete time intervals the optimal weights can differ markedly from the ExpS weights. We show that, conditional on the break size and date, the optimal weights follow a step function that allocates constant weights within regimes but different weights between regimes. A striking result emerges under multiple breaks: observations of the last regime that continues into the forecast period may not receive the highest weight. The intuition for this result is that the bias component of the MSFE can be reduced by giving the largest weights to observations in an earlier regime to counterbalance biases of the opposite sign in another regime.

In practice, dates and sizes of the breaks are unknown and must be estimated. As such estimates tend to be quite imprecise and their use in practice leads to a deterioration of forecasts, which can be quite substantial. In order to address this problem, we develop weights that are robust to the uncertainty that surrounds the dates and the sizes of the breaks. Robust optimal weights are derived by integrating the optimal weights with respect to uniformly distributed break dates. An interesting insight from these derivations is that the effect of uncertainty of the break size on the

[1] There are many statistical procedures that can be used for detection of break dates, such as Brown et al. (1975), Andrews (1993), Andrews et al. (1996), Bai and Perron (1997, 2003), and Altissimo and Corradi (2003).

weights is of order T^{-2} if the break is in the slope coefficient, and of order T^{-3} if the break is in the error variances, where T is the full sample size that include the pre-break observations. In contrast, the uncertainty around the break date is of order T^{-1}, which suggests that dating a break correctly is generally more important than knowing the precise size of the break.

We conduct Monte Carlo experiments that compare the forecasts from optimal and robust optimal weights to a range of alternative forecasting methods. It emerges that the key factor for the relative performance of different forecasting methods under a discrete break is the size of the break. A larger break leads to more precise estimates of the break date and improves forecasts that are conditional on these estimates, which include the optimal weights forecast, post-break forecasts, and optimal window forecasts. In contrast, when the break is small relative to the noise in the DGP, the robust optimal weights produce the best forecasts as they do not make use of the often imprecisely estimated break dates and sizes. When the break process is continuous, ExpS forecasts that estimate the down-weighting parameter perform well. However, even under the continuous break process the forecasts from the robust optimal weights perform well and in some settings provide the best forecasts.

We use the different methods considered in this paper to forecast real GDP using the slope of the yield curve across nine industrial economies over the period 1994Q1–2009Q4. The general finding is that breaks are difficult to estimate with sufficient accuracy and, similar to the Monte Carlo results, forecasts based on estimates of break dates perform poorly. Forecasts based on robust optimal weights deliver the largest improvements over forecasts based on equal weights, and these improvements are statistically significant.

The rest of the paper is set out as follows. Using a linear regression model, derivations of optimal weights under different break processes are set out in Section 2, and the MSFE outcomes are compared across different forecasting methods. Optimal weights that are robust to the uncertainty of the break process are motivated and derived in Section 3. Monte Carlo evidence on the comparative performance of the different forecasting methods is discussed in Section 4. Empirical results are presented in Section 5. The paper ends with some concluding remarks in Section 6. A few of the less essential derivations are collected in a mathematical appendix. Additional material can be found in a web supplement.

2 Optimal weights under different break processes

Consider the linear regression model

$$y_t = \boldsymbol{\beta}_t' \mathbf{x}_t + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim iid(0,1), \quad t = 1, 2, \ldots, T, T+1 \tag{1}$$

where $\mathbf{x}_t$ is a $k \times 1$ vector of stationary regressors, and the $k \times 1$ coefficient vector, $\boldsymbol{\beta}_t$, and the scalar error variance, σ_t^2, are subject to breaks. We consider two possible types of break processes. A continuous break process whereby $\boldsymbol{\beta}_t$ changes in every period by a relatively small amount. A prominent example is the random walk model

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{S}_\beta \mathbf{v}_t, \quad \text{where } \mathbf{v}_t \sim iid(\mathbf{0}, \boldsymbol{I}_k),$$

where $\boldsymbol{I}_k$ is the identity matrix of order k, and the break variance, $\boldsymbol{\Sigma}_\beta = \boldsymbol{S}_\beta \boldsymbol{S}_\beta'$, is assumed to be small relative to σ_t^2.[2] Additionally, σ_t may be subject to a similar break

[2]The covariance matrix $\boldsymbol{\Sigma}_\beta$ is said to be small relative to σ_t if $\|\boldsymbol{\Sigma}_\beta\| / \sigma_t$ is small, where $\|\boldsymbol{A}\|^2 = \text{tr}(\mathbf{A}\mathbf{A}')$ denotes the Euclidean norm of matrix $\mathbf{A}$.

process. Alternatively, the breaks could be discrete where the parameters change at a small number of distinct points in time, $T_{b,i}$, $i = 1, 2, \ldots, n$,[3]

$$\boldsymbol{\beta}_t = \begin{cases} \boldsymbol{\beta}_{(1)} & \text{for } 1 < t \leq T_{b,1} \\ \boldsymbol{\beta}_{(2)} & \text{for } T_{b,1} < t \leq T_{b,2} \\ \vdots & \\ \boldsymbol{\beta}_{(n+1)} & \text{for } T_{b,n} < t \leq T \end{cases}$$

In contrast to the continuously changing parameter model, the number of discrete breaks, n, is assumed to be small, although the break sizes, measured by $\left\|\boldsymbol{\beta}_{(i)} - \boldsymbol{\beta}_{(i-1)}\right\|$ could be large relative to σ_t. There are merits in both specifications, and a choice between them would depend on the particular data at hand.

We propose a general approach to achieve a minimum mean square forecast error (MSFE) under both break processes. We weight past observations by weights w_t in the estimation

$$\hat{\boldsymbol{\beta}}_T(\mathbf{w}) = \left(\sum_{t=1}^{T} w_t \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \sum_{t=1}^{T} w_t \mathbf{x}_t \mathbf{y}_t,$$

subject to the restriction $\sum_{t=1}^{T} w_t = 1$. The weights $\mathbf{w} = (w_1, w_2, \ldots, w_T)'$ are chosen such that the resulting MSFE of the one-step ahead forecast, $\hat{y}_{T+1} = \hat{\boldsymbol{\beta}}_T' \mathbf{x}_{T+1}$, is minimized.

Closed form solutions under the continuous break process are only available when we simplify the model to one without time-varying regressors. In this setting the optimal weights recover the exponential smoothing forecast. For the discrete break process we derive new results for the same simple model but also for models with one or more regressors.

2.1 Optimal weights in a model with continuous breaks

Consider the following model

$$y_t = \beta_t + \sigma_\varepsilon \varepsilon_t, \tag{2}$$

where $\beta_t = \beta_{t-1} + \sigma_v v_t$, and ε_t and v_t are $iid(0, 1)$. The optimal weights for a one-step ahead forecast can be found by minimizing $\mathrm{E}(y_{T+1} - \Sigma_{t=1}^{T} w_t y_t)^2$ with respect to w_t, $t = 1, 2, \ldots, T$, subject to $\sum_{t=1}^{T} w_t = 1$. For a solution to this problem we first note that the forecast error is given by

$$e_{T+1} = y_{T+1} - \hat{\beta}_{T+1}(\mathbf{w}) = \beta_{T+1} - \mathbf{w}'\boldsymbol{\beta} + \sigma_\varepsilon(\varepsilon_{T+1} - \mathbf{w}'\boldsymbol{\varepsilon}),$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_T)'$. But using the random walk formulation of $\boldsymbol{\beta}$ we have $\boldsymbol{\beta} = \beta_0 \boldsymbol{\iota}_T + \sigma_v \mathbf{H}\mathbf{v}$, where $\mathbf{v} = (v_1, v_2, \ldots, v_T)'$,

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}, \text{ and } \boldsymbol{\iota}_T = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}.$$

[3]Note that parentheses around subscripts denote subsamples between breaks, such that β_t is the parameter at period t but $\beta_{(i+1)}$ the parameter after break i.

Also, $\beta_{T+1} = \beta_0 + \sigma_v \boldsymbol{\iota}_T' \mathbf{v} + \sigma_v v_{T+1}$. Hence, $\sigma_\varepsilon^{-1} e_{T+1} = (\boldsymbol{\iota}_T' \mathbf{v} - \mathbf{w}'\mathbf{H}\mathbf{v})\,\delta + (\varepsilon_{T+1} - \mathbf{w}'\boldsymbol{\varepsilon}) + \delta v_{T+1}$, where $\delta^2 = \sigma_v^2/\sigma_\varepsilon^2$. Therefore, (noting that by assumption $\mathbf{v}$ and $\boldsymbol{\varepsilon}$ are independently distributed)

$$\mathrm{E}(\sigma_\varepsilon^{-2} e_{T+1}^2 \,|\mathbf{w}\,) \propto \delta^2 \mathbf{w}'\mathbf{H}\mathbf{H}'\mathbf{w} - 2\delta^2 \mathbf{w}'\mathbf{H}\boldsymbol{\iota}_T + \mathbf{w}'\mathbf{w}.$$

The first order condition for minimization of $\mathrm{E}(\sigma_\varepsilon^{-2} e_{T+1}^2 \,|\mathbf{w}\,)$ subject to the constraint, $\mathbf{w}'\boldsymbol{\iota}_T = 1$, is given by $\delta^2 \mathbf{H}\mathbf{H}'\mathbf{w} - \delta^2 \mathbf{H}\boldsymbol{\iota}_T + \mathbf{w} - \theta\boldsymbol{\iota}_T = 0$, where θ is the Lagrangian multiplier associated with $\mathbf{w}'\boldsymbol{\iota}_T = 1$. Solving for the optimal weights, $\mathbf{w}^*(\delta)$, in terms of θ we have

$$\mathbf{w}^*(\delta) = (\delta^2 \mathbf{H}\mathbf{H}' + \mathbf{I}_T)^{-1}(\delta^2 \mathbf{H} + \theta \mathbf{I}_T)\boldsymbol{\iota}_T, \tag{3}$$

Also, since $\boldsymbol{\iota}_T'\mathbf{w} = 1$,

$$\theta = \frac{1 - \boldsymbol{\iota}_T'(\delta^2 \mathbf{H}\mathbf{H}' + \mathbf{I}_T)^{-1}\delta^2 \mathbf{H}\boldsymbol{\iota}_T}{\boldsymbol{\iota}_T'(\delta^2 \mathbf{H}\mathbf{H}' + \mathbf{I}_T)^{-1}\boldsymbol{\iota}_T}. \tag{4}$$

For the extreme values of $\delta^2 = \infty$ and 0 we obtain the random walk and equal weighted solutions: $\mathbf{w}^*(\infty) = (1, 0, \ldots, 0)'$ and $\mathbf{w}^*(0) = T^{-1}(1, 1, \ldots, 1)'$, respectively.[4]

The literature on exponential smoothing has traditionally used a different solution to address the time varying β_t. Write the model in terms of the observables

$$y_t - y_{t-1} = \sigma_v v_t + \sigma_\varepsilon(\varepsilon_t - \varepsilon_{t-1}), \tag{5}$$

which represents an MA(1) process in Δy_t with the MA parameter given by γ. More specifically

$$\Delta y_t = \xi_t - \gamma \xi_{t-1}, \tag{6}$$

where ξ_t is a serially uncorrelated process with mean zero and a constant variance, and

$$\frac{\gamma}{1+\gamma^2} = \frac{\sigma_\varepsilon^2}{2\sigma_\varepsilon^2 + \sigma_v^2} = \frac{1}{2+\delta^2}.$$

Hence,

$$\gamma^2 - (2+\delta^2)\gamma + 1 = 0. \tag{7}$$

or

$$\delta = (1-\gamma)/\sqrt{\gamma}. \tag{8}$$

To solve for γ, note that (7) has two real roots given by

$$\gamma = \frac{(2+\delta^2) \pm \delta(4+\delta^2)^{1/2}}{2}. \tag{9}$$

Since $\delta > 0$, then $\gamma = 1 + \delta^2/2 - \delta(1 + \delta^2/4)^{1/2}$ is the root that lies within the unit circle and should be used.[5] The optimal forecast of y_{T+1} is now given by

$$\mathrm{E}(y_{T+1} \,|y_T, y_{T-1}, \ldots) = y_T - \gamma \xi_T,$$

but since $0 < \gamma < 1$ we can invert the MA process to obtain $\xi_T = (1 - \gamma L)^{-1}(y_T - y_{T-1})$, and

$$\begin{aligned} \mathrm{E}(y_{T+1} \,|y_T, y_{T-1}, \ldots) &= y_T - \gamma(1-\gamma L)^{-1}(y_T - y_{T-1}), \\ &= (1-\gamma)\left(y_T + \gamma y_{T-1} + \gamma^2 y_{T-2} + \ldots\right). \end{aligned}$$

[4] Derivations of these results are available in web supplement B.1.

[5] Since $\delta > 0$ then it is easily seen that $0 < \gamma = 1 + \delta^2/2 - \delta(1 + \delta^2/4)^{1/2} < 1$.

In practice, the infinite series must be truncated to yield the ExpS forecast

$$\hat{y}_{T+1} = \frac{1-\gamma}{1-\gamma^T}\sum_{j=1}^{T}\gamma^{T-j}y_j = \sum_{j=1}^{T} w_j^{(e)}(\gamma)y_j \tag{10}$$

and the quality of the approximation will depend on T and γ, and could be poor when T is relatively small and γ close to unity. For large T and γ not too close to unity, the elements of $w_t^{(e)}$ will be very close to $(1-\gamma)\gamma^{T-t}$ for $t = T, T-1, \ldots$ It is worth noting that the weights $w_t^{(e)}$ add up to unity and adapt to the sample size T, whilst the MA weights (γ^{T-j}) may not when γ is not too close to unity.

The MA weights can be viewed as an approximation to the optimal weights, $\mathbf{w}^*(\delta)$ given by (3), when T is sufficiently large.

2.2 Optimal weights in a model with a single, discrete break

Again consider model (2) but now assume that β_t is subject to a single, discrete break at T_b, $1 < T_b < T$,

$$\beta_t = \begin{cases} \beta_{(1)} & \text{for } t \le T_b \\ \beta_{(2)} & \text{for } T_b < t \le T+1 \end{cases}$$

In this case the forecast is $\hat{y}_{T+1} = \hat{\beta}_T(\mathbf{w})$ where $\hat{\beta}_T(\mathbf{w}) = \sum_{t=1}^{T} w_t y_t$ and

$$\hat{\beta}_T(\mathbf{w}) - \beta_T = (\beta_{(1)} - \beta_{(2)})\sum_{t=1}^{T_b} w_t + \sigma_\varepsilon \sum_{t=1}^{T} w_t \varepsilon_t.$$

Therefore, the forecast error is given by

$$e_{T+1}(\mathbf{w}) = y_{T+1} - \hat{\beta}_T(\mathbf{w}) = \sigma_\varepsilon \varepsilon_{T+1} - (\beta_{(1)} - \beta_{(2)})\sum_{t=1}^{T_b} w_t - \sigma_\varepsilon \sum_{t=1}^{T} w_t \varepsilon_t,$$

and the MSFE scaled by the error variance is

$$\mathrm{E}[\sigma_\varepsilon^{-2} e_{T+1}^2(\mathbf{w})] = 1 + \lambda^2 \left(\sum_{t=1}^{T_b} w_t\right)^2 + \sum_{t=1}^{T} w_t^2, \tag{11}$$

where $\lambda = (\beta_{(1)} - \beta_{(2)})/\sigma_\varepsilon$.

We can now obtain the optimal weights by minimizing (11) subject to $\sum_{t=1}^{T} w_t = 1$. The first order conditions are: $2\lambda^2 \sum_{t=1}^{T_b} w_t + 2w_t + \theta = 0$ for $t \le T_b$, and $2w_t + \theta = 0$ for $T_b < t \le T$, where θ is the Lagrange multiplier associated with $\sum_{t=1}^{T} w_t = 1$. Note that w_t for $t \le T_b$ does not depend on t. The same is also true of w_t for $t > T_b$. Hence, defining the weights for each pre-break observation as $w_{(1)}$ and those for each post-break observation as $w_{(2)}$, we obtain

$$w_t = \begin{cases} w_{(1)} = -\lambda^2 \sum_{t=1}^{T_b} w_t - \theta/2 & \text{for } 1 < t \le T_b \\ w_{(2)} = -\theta/2 & \text{for } T_b < t \le T \end{cases}$$

and $w_{(2)} - w_{(1)} = \lambda^2 \sum_{t=1}^{T_b} w_t = \lambda^2 T_b w_{(1)}$. Solving for $w_{(2)}$ and substituting into $\sum_{t=1}^{T} w_t = T_b w_{(1)} + (T - T_b) w_{(2)} = 1$ now yields the optimal weights

$$w_{(1)} = \frac{1}{T}\frac{1}{1 + Tb(1-b)\lambda^2}, \tag{12}$$

$$w_{(2)} = \frac{1}{T}\frac{1 + Tb\lambda^2}{1 + Tb(1-b)\lambda^2}. \tag{13}$$

where $b = T_b/T$.

We can use the fact that the weights are constant in the sub-samples in (11) to obtain the scaled MSFE: $\mathrm{E}(\sigma_\varepsilon^{-2} e_{T+1}^2 \left| w_{(1)}, w_{(2)} \right.) = 1+(T_b\lambda w_{(1)})^2+T_b w_{(1)}^2+(T-T_b)w_{(2)}^2$ and using (12) and (13) it is straightforward to show that the above result reduces to

$$\mathrm{E}(\sigma_\varepsilon^{-2} e_{T+1}^2 \left| w_{(1)}, w_{(2)} \right.) = 1 + \frac{1}{T}\frac{1+Tb\lambda^2}{1+Tb(1-b)\lambda^2} = 1 + w_{(2)}. \tag{14}$$

Namely, the MSFE varies with λ through the post-beak weight, $w_{(2)}$.

We can now compare the forecasts based on optimal weights to those from a range of alternative forecasting methods: post-break window observations, the optimal estimation window, averaging across estimation windows, and exponential smoothing.

2.2.1 Optimal window and post-break window forecasts

The optimal window choice proposed in Pesaran and Timmermann (2007), gives equal weights to observations within the window and zero weights to preceding observations. Suppose that the optimal window size contains observations T_v to T (inclusive), where $v = (T - T_v + 1)/T$ so that $T_v = T(1-v)+1$. Then, the optimal window size is[6]

$$v^o = \begin{cases} \frac{1-b}{1-\frac{1}{2\lambda^2(1-b)T}} & \text{if } \lambda^2 \geq \frac{T}{2(T-T_b)T_b} \\ 1 & \text{if } \lambda^2 < \frac{T}{2(T-T_b)T_b}. \end{cases}$$

The scaled MSFE for the optimal window is (for $\lambda^2 \geq \frac{T}{2(T-T_b)T_b}$)

$$\mathrm{E}\left(\sigma_\varepsilon^{-2} e_{T+1}^2 | v^o_{v>(1-b)}\right) = 1 + \frac{1}{T(1-b)} - \frac{1}{T^2}\frac{1}{4\lambda^2(1-b)^2}. \tag{15}$$

Furthermore, the scaled MSFE of the post-break window is

$$\mathrm{E}\left[\sigma_\varepsilon^{-2} e_{T+1}^2 | v = (1-b)\right] = 1 + \frac{1}{T(1-b)}, \tag{16}$$

and it can be seen that the MSFE in (15) cannot be greater than that in (16) as the last term of (15) is non-negative.

In order to compare the MSFEs of the forecasts from the optimal window to that of the optimal weights forecast, we can use (15) and (14), which yields

$$\begin{aligned} &\mathrm{E}\left(\sigma_\varepsilon^{-2} e_{T+1}^2 | v^o_{v>(1-b)}\right) - \mathrm{E}(\sigma_\varepsilon^{-2} e_{T+1}^2 \left| w_{(1)}, w_{(2)} \right.) \\ &= \left[\frac{1}{T(1-b)} - \frac{1}{T^2}\frac{1}{4\lambda^2(1-b)^2}\right] - \frac{1}{T}\frac{1+Tb\lambda^2}{1+Tb(1-b)\lambda^2}, \\ &= \frac{1}{T}\frac{T\lambda^2 b(1-b) + 2T\lambda^2 b(1-b) - 1}{4T(1-b)^2\lambda^2[1+Tb(1-b)\lambda^2]} > 0, \end{aligned} \tag{17}$$

where the last inequality follows since $v^o \leq 1$, implies that $T\lambda^2 b(1-b) \geq 1/2$. Therefore, forecasts obtained from optimal weights will have a smaller MSFE than forecasts giving equal weight to observations in an optimally chosen window size. In the case where $T\lambda^2 < 1/2$, the optimal window contains all observations, so that the comparison

[6] Derivations can be found in web supplement B.2.

Table 1: Relative MSFE for a single break in drift for known b and λ

b	0.95			0.9		
λ	0.5	1	2	0.5	1	2
opt. weights	0.901	0.610	0.258	0.884	0.600	0.258
post-break obs.	0.971	0.628	0.260	0.907	0.604	0.259
opt. window	0.939	0.622	0.259	0.899	0.603	0.259
AveW($v_{\min} = 0.05$)	0.966	0.900	0.829	0.941	0.830	0.704
ExpS($\gamma = 0.95$)	0.973	0.924	0.872	0.958	0.883	0.799

Note: The table reports the MSFE relative to that of the equal weights forecasts, $\text{MSFE}_i/\text{MSFE}_{\text{equal}}$, where MSFE_i is the MSFE of the respective forecasting method in the first column. These are (i) using the optimal weights, (ii) using the post-break observations, (iii) forecasts based on the optimal window, (iv) AveW forecasts with $v_{\min} = 0.05$ and $m = T(1 - v_{\min}) + 1$ windows, and (v) ExpS forecasts with $\gamma = 0.95$. Finally, $T = 100$.

is between the optimal weights and equal weights. Clearly, by merit of the optimality of the weights the forecast based on optimal weights will have a lower MSFE.

While it is clear from the above that using optimal weights decreases the MSFE, it is interesting to get a quantitative sense of the difference in MSFEs. For a range of values for λ and b, Table 1 shows the relative performance of different forecasting methods. That is, for forecasting method i we report $\text{MSFE}_i/\text{MSFE}_{equal}$, where subscript *equal* denotes equal weights forecasts. The first line gives the relative performance for optimal weights, the second line for the forecasts based on post-break window, and the third line gives the results for the optimal window.

It can be seen that the forecasts based on optimal weights has the lowest MSFE across all parameter combinations. The MSFE of the forecasts based on the post-break window is similar to that using optimal weights when either the break or the post-break window is large ($b = 0.9$). For breaks of smaller magnitude, however, post-break window forecasts have substantially larger MSFEs. Forecasts based on the optimal windows perform better than those based on post-break windows but for small breaks have substantially larger MSFEs than the optimal weights forecasts.

2.2.2 Averaging across estimation windows

Pesaran and Pick (2011) discuss theoretical properties of averaging forecasts across estimation windows (AveW). For the random walk (2), the AveW forecast is

$$\hat{y}_{T+1} = m^{-1} \sum_{i=1}^{m} \hat{y}_{T+1}(v_{(i)}), \quad \text{where} \quad \hat{y}_{T+1}(v_{(i)}) = \frac{1}{T - T_{v_{(i)}} + 1} \sum_{s=T_{v_{(i)}}}^{T} y_s,$$

$v_{(i)}$ is the minimum (shortest) window, and m is the number of estimation windows. The AveW forecast has the MSFE

$$\text{E}(\sigma_\varepsilon^{-2} e_{T+1}^2 | v_{(i)}) = 1 + \left[\frac{\lambda}{m} \sum_{i=1}^{m} \frac{v_{(i)} - (1-b)}{v_{(i)}} \text{I}[v_{(i)} - (1-b)] \right]^2 + \frac{1}{m^2} \sum_{i=1}^{m} \frac{1 + 2(i-1)}{T v_{(i)}}.$$

Pesaran and Pick (2011) show that for the case of the random walk it will improve over equal weights forecasts using all observations unless the break is very small. This

is reflected in the results for $v_{\min} = 0.05$ in the fourth line of Table 1. The AveW forecasts have smaller MSFEs than the single window forecasts using equal weights but they have substantially larger MSFEs than the forecasts obtained using the optimal weights. The intuition for this result is that averaging over estimation windows can be seen as weighting observations with decaying weights. The optimal weights (12) and (13), however, have a discrete change and will only be approximated poorly by the weights implied by the AveW forecast. Given the optimality of $w_{(1)}$ and $w_{(2)}$, the AveW MSFEs will necessarily be larger than those of the optimal weights forecasts. However, these results are not surprising as averaging forecasts is based on the idea that it will be beneficial when the break date and size are uncertain. We will explore such settings in the Monte Carlo experiments in Section 4.

2.2.3 Exponential smoothing

In Section 2.1 we have shown that under continuous breaks optimal weights recover ExpS weights for large T or a downweighting parameter not too close to unity. While the application of ExpS weights is not optimal under discrete breaks, it is nevertheless interesting to get a quantitative measure of the loss implied in using weights for continuous breaks when there is a single discrete break.

The MSFE of the exponential smoothing forecast is (Pesaran and Pick 2011)

$$\mathrm{E}(\sigma_{\varepsilon}^{-2} e_{T+1}^2 | \gamma) = 1 + \lambda^2 \left(\frac{\gamma^{1+Tb} - \gamma^T}{1 - \gamma^T} \right)^2 + \left(\frac{1-\gamma}{1-\gamma^T} \right) \left(\frac{1-\gamma^{2T}}{1-\gamma^2} \right).$$

The last line in Table 1 reports results for $\gamma = 0.95$. Similar to the AveW forecasts, the ExpS forecasts improve on the results from the forecasts using equal weights but have a larger MSFE than the forecasts based on the optimal weights. The reason is that, just as the AveW forecasts, the ExpS forecasts use smoothly decaying weights for the observations, whilst it has been shown that in the presence of discrete breaks, discretely changing weights are optimal.

2.3 A single, discrete break in a multiple regression model

We now turn to the multiple regression model where the slope parameters and the error variance are subject to a single break at time $t = T_b$

$$y_t = \begin{cases} \boldsymbol{\beta}_{(1)}' \mathbf{x}_t + \sigma_{(1)} \varepsilon_t & \text{for } 1 \leq t \leq T_b, \\ \boldsymbol{\beta}_{(2)}' \mathbf{x}_t + \sigma_{(2)} \varepsilon_t & \text{for } T_b + 1 \leq t \leq T \end{cases}, \tag{18}$$

where $\mathbf{x}_t$ is a $k \times 1$ vector of exogenous regressors and $\varepsilon_t \sim iid(0,1)$. Again, suppose that the slope parameter is estimated by weighting observations over the whole sample

$$\hat{\boldsymbol{\beta}}_T(\mathbf{w}) = \left(\sum_{t=1}^{T} w_t \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^{T} w_t \mathbf{x}_t y_t. \tag{19}$$

The scaled MSFE is

$$\begin{aligned} &\mathrm{E}\left[\sigma_{(2)}^{-2} e_{T+1}^2(\mathbf{w}) \,|\mathbf{x}_t,\ t = 1, 2, \ldots, T+1 \right] \\ &\quad = 1 + \left[\mathbf{x}_{T+1}' \mathbf{S}^{-1}(\mathbf{w}) \mathbf{S}_1(\mathbf{w}_{(1)}) \boldsymbol{\lambda} \right]^2 \\ &\qquad + \mathbf{x}_{T+1}' \mathbf{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T_b} q^2 w_t^2 \mathbf{x}_t \mathbf{x}_t' + \sum_{t=T_b+1}^{T} w_t^2 \mathbf{x}_t \mathbf{x}_t' \right) \mathbf{S}^{-1}(\mathbf{w}) \mathbf{x}_{T+1}, \end{aligned} \tag{20}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(2)})/\sigma_{(2)}$, $q = \sigma_{(1)}/\sigma_{(2)}$, $\mathbf{S}(\mathbf{w}) = \mathbf{S}_1(\mathbf{w}_{(1)}) + \mathbf{S}_2(\mathbf{w}_{(2)})$, $\mathbf{S}_1(\mathbf{w}_{(1)}) = \sum_{t=1}^{T_b} w_t \mathbf{x}_t \mathbf{x}_t'$, and $\mathbf{S}_2(\mathbf{w}_{(2)}) = \sum_{t=T_b+1}^{T} w_t \boldsymbol{x}_t \mathbf{x}_t'$. See Appendix A.1 for details.

Similar to the derivations in Section 2.2, minimizing (20) yields the optimal weights. For $t \leq T_b$ we have

$$\begin{aligned}\left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t\right] q^2 w_t &= \mathbf{x}_{T+1}'\boldsymbol{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T_b} q^2 w_t^2 \mathbf{x}_t \mathbf{x}_t' + \sum_{t=T_b+1}^{T} w_t^2 \mathbf{x}_t \mathbf{x}_t' \right) \mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t \\ &\quad \left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right] \left[\mathbf{x}_t'\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}\right], \qquad (21)\end{aligned}$$

and for $t \geq T_b + 1$

$$\begin{aligned}\left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t\right] w_t &= \mathbf{x}_{T+1}'\boldsymbol{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T_b} q^2 w_t^2 \mathbf{x}_t \mathbf{x}_t' + \sum_{t=T_b+1}^{T} w_t^2 \mathbf{x}_t \mathbf{x}_t' \right) \mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t \\ &\quad + \left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right] \left[\mathbf{x}_t'\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right], \qquad (22)\end{aligned}$$

with the details provided in Appendix A.1.

These optimal weights have a number of interesting properties. First, in the absence of a break, that is when $\boldsymbol{\lambda} = \mathbf{0}$ and $q = 1$, then $w_t = w$ for all t, as to be expected. To see this, note that when $\boldsymbol{\lambda} = \mathbf{0}$ and $q = 1$, then for all t we have

$$w_t = \frac{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T} w_t^2 \mathbf{x}_t \mathbf{x}_t' \right) \mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t}{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t}.$$

It is now easily seen that $w_t = w$ (fixed) is a solution to the above. Note that for $w_t = w$, we have $\mathbf{S}(\mathbf{w}) = w\mathbf{S}(1)$ and therefore

$$w_t = \frac{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(1) \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right) \mathbf{S}^{-1}(1)\mathbf{x}_t}{w^{-1}\mathbf{x}_{T+1}'\mathbf{S}^{-1}(1)\mathbf{x}_t} = \frac{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(1)\mathbf{S}(1)\mathbf{S}^{-1}(1)\mathbf{x}_t}{w^{-1}\mathbf{x}_{T+1}'\mathbf{S}^{-1}(1)\mathbf{x}_t} = w.$$

Consider now the case where $\boldsymbol{\lambda} \neq \mathbf{0}$ and $q \neq 1$, but suppose that $\mathbf{x}_1 = \mathbf{x}_2$. Then, using (21) and (22), we have that for the optimal weights for $t = 1$ and $t = 2$

$$q^2 \left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_1\right] (w_2 - w_1) = 0.$$

Hence, the weights within a given regime will be the same if the regressor values are identical. But the same is not true of the weights for time points in different regimes. For example, for the first regime select $t = 1$ and for the second regime select $t = T$, and suppose that $\mathbf{x}_1 = \mathbf{x}_T$. Then from (21) and (22), and recalling that $\mathbf{S}_1(\mathbf{w}_{(1)}) + \mathbf{S}_2(\mathbf{w}_{(2)}) = \mathbf{S}(\mathbf{w})$, we have

$$\left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_T\right] (w_T - q^2 w_1) = \left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right] \left(\mathbf{x}_T'\boldsymbol{\lambda}\right),$$

which suggests that when $\boldsymbol{\lambda} \neq \mathbf{0}$ and $q \neq 1$, the weights across the two regimes differ even if the regressor values are the same. Therefore, in general, the optimal weights will differ both within and across regimes.

An exact analytical solution does not seem to be available. However, analytic solutions can be derived if $k = 1$ or asymptotically when T is sufficiently large. In general, the unknown weights in (21) and (22) must be solved numerically. Some notes on the implementation of a suitable numerical procedure are provided in web supplement B.3.

2.3.1 A single, discrete break in a model with one regressor

In the case where $k = 1$, the scaled MSFE (20) simplifies to

$$\mathrm{E}[\sigma_{(2)}^{-2} e_{T+1}^2(\mathbf{w})] = 1 + \left[\frac{x_{T+1} S(\mathbf{w}_{(1)})\lambda}{S(\mathbf{w})}\right]^2 + \frac{x_{T+1}^2 \left(\sum_{t=1}^{T_b} q^2 w_t^2 x_t^2 + \sum_{t=T_b+1}^{T} w_t^2 x_t^2\right)}{[S(\mathbf{w})]^2}, \tag{23}$$

and the first order conditions (21) and (22) reduce to

$$w_t = \begin{cases} \frac{\sum_{t=1}^{T} w_t^2 x_t^2}{q^2 S(\mathbf{w})} - \lambda^2 \frac{S_1(\mathbf{w}_{(1)}) S_2(\mathbf{w}_{(2)})}{q^2 S(\mathbf{w})} & \text{for } t \leq T_b, \\ \frac{\sum_{t=1}^{T} w_t^2 x_t^2}{S(\mathbf{w})} + \lambda^2 \frac{S_1^2(\mathbf{w}_{(1)})}{S(\mathbf{w})} & \text{for } t \geq T_b + 1. \end{cases}$$

Similar to the case of model (2), w_t for $t \leq T_b$ does not depend on t and the same is true for w_t for $t > T_b$. We can therefore again denote the pre-break weights by $w_{(1)}$ and the post-break weights by $w_{(2)}$. Using the above results it now readily follows that $w_{(2)} - q^2 w_{(1)} = w_{(1)} S_1(1)\lambda^2$, where $S_1(1) = \sum_{t=1}^{T_b} x_t^2$. Also using the constraint $\sum_{t=1}^{T} w_t = 1$ we have $w_{(1)} T_b + (T - T_b) w_{(2)} = 1$. Hence, for T_b reasonably large, which is not a restrictive assumption for the problem under consideration, and solving for $w_{(1)}$ and $w_{(2)}$ we obtain

$$w_{(1)} = \frac{1}{T} \frac{1}{b + (1-b)(q^2 + Tb\lambda^2 \omega_x^2)}, \tag{24}$$

$$w_{(2)} = \frac{1}{T} \frac{q^2 + Tb\lambda^2 \omega_x^2}{b + (1-b)(q^2 + Tb\lambda^2 \omega_x^2)}, \tag{25}$$

where $\omega_x^2 = \mathrm{plim}_{T_b \to \infty}(\frac{1}{T_b} \sum_{t=1}^{T_b} x_t^2)$.

Given that for the optimal weights $w_t^{(\mathrm{opt.})} = w_{(1)}$ for $t \leq T_b$ and $w_t^{(\mathrm{opt.})} = w_{(2)}$ for $t > T_b$, (23) can be rewritten as

$$E\left(\sigma_{(2)}^{-2} e_{T+1}^2 | w_t^{(\mathrm{opt.})}\right) \approx 1 + \frac{x_{T+1}^2}{T\omega_x^2} \frac{w_{(1)}^2 b(Tb\phi^2 + q^2) + w_{(2)}^2 (1-b)}{\left[w_{(1)} b + (1-b) w_{(2)}\right]^2}. \tag{26}$$

From (24) and (25) it can be seen that $w_{(2)} = (Tb\phi^2 + q^2) w_{(1)}$ and (26) simplifies to

$$E\left(\sigma_{(2)}^{-2} e_{T+1}^2 | w_t^{(\mathrm{opt.})}\right) \approx 1 + \frac{x_{T+1}^2}{\omega_x^2} w_{(2)}.$$

Namely, the MSFE varies with λ and q through the post-break weight, $w_{(2)}$.

In the standard case where all observations are given equal weights, namely $w_t^{(\mathrm{equal})} = 1/T$, we have

$$\begin{aligned} E\left(\sigma_{(2)}^{-2} e_{T+1}^2 | w_t^{(\mathrm{equal})}\right) &= 1 + b^2 \phi^2 \frac{x_{T+1}^2}{\omega_x^2} \left(\frac{T_b^{-1} \sum_{t=1}^{T_b} x_t^2}{T^{-1} \sum_{t=1}^{T} x_t^2} \right)^2 \\ &\quad + \frac{b\left(q^2 - 1\right) x_{T+1}^2 \left(T_b^{-1} \sum_{t=1}^{T_b} x_t^2 \right)}{\left(T^{-1} \sum_{t=1}^{T} x_t^2 \right)^2} + \frac{1}{T} \left(\frac{x_{T+1}^2}{T^{-1} \sum_{t=1}^{T} x_t^2} \right). \end{aligned}$$

The parameters ϕ and q measure the sizes of the breaks in β and σ, and $b = T_b/T$ gives the proportion of pre-break observations.

For T_b sufficiently large and $T - T_b$ small, we have the approximation

$$\mathrm{E}\left(\sigma_{(2)}^{-2}e_{T+1}^2|w_t^{(\text{equal})}\right) \approx 1 + \frac{x_{T+1}^2}{\omega_x^2}\left[b^2\phi^2 + \frac{bq^2 + (1-b)}{T}\right].$$

Comparing the MSFE of optimal weights with the one based on equal weights we have

$$\begin{aligned}\frac{\omega_x^2}{x_{T+1}^2}\left(\mathrm{MSFE}_{\text{equal}} - \mathrm{MSFE}_{\text{opt.}}\right) &= \frac{b\left(Tb\phi^2 + q^2\right) + (1-b)}{T} - \frac{Tb\phi^2 + q^2}{T\left[b + (1-b)(Tb\phi^2 + q^2)\right]} \\ &= \frac{b(1-b)[1 - Tb\phi^2 - q^2]^2}{T\left[b + (1-b)(Tb\phi^2 + q^2)\right]} > 0.\end{aligned}$$

Similarly, when w_t are set independently of x_t (as in the case of exponential down-weighting) we have

$$E\left(\sigma_{(2)}^{-2}e_{T+1}^2\right) \approx 1 + \frac{x_{T+1}^2}{\omega_x^2}\left[\phi^2\left(\sum_{t=1}^{T_b} w_t\right)^2 + \left(q^2 - 1\right)\sum_{t=1}^{T_b} w_t^2 + \sum_{t=1}^{T} w_t^2\right].$$

When only post-break observations are used, the implicit weights are $w_t^{(\text{post})} = 0$ for $t \leq T_b$ and $w_t^{(\text{post})} = (T - T_b)/T$ for $t > T_b$. We therefore have

$$\mathrm{E}\left(\sigma_{(2)}^{-2}e_{T+1}^2|w_t^{(\text{post})}\right) \approx 1 + \frac{x_{T+1}^2}{\omega_x^2}\frac{1}{T(1-b)}.$$

Comparing this to the MSFE based on optimal weights we have

$$\frac{\omega_x^2}{x_{T+1}^2}\left(\mathrm{MSFE}_{\text{post}} - \mathrm{MSFE}_{\text{opt.}}\right) = \frac{b}{T\left(1-b\right)\left[b + (1-b)(Tb\phi^2 + q^2)\right]} > 0,$$

namely, optimal weights forecasts dominate post-break forecasts for all values of $0 < b < 1$, but, as to be expected, the superiority of the optimal weights forecasts diminishes as $T(1-b) \to \infty$.

2.3.2 Asymptotic results with $k \geq 1$ stationary regressors

The general solution in (21) and (22) can be simplified if we assume that T and T_b are sufficiently large with $T - T_b$ fixed, and $\mathbf{x}_t$ is a stationary process with $\mathrm{E}(\mathbf{x}_t\mathbf{x}_t') = \mathbf{\Omega}_{xx}$ a positive definite matrix. That is we assume that $T \to \infty$ and $b \to 1$ but $T(1-b) \to \tau$, where τ is a relatively small, constant number of post-break observations. Under these assumptions (and conditional on the weights, w_t)

$$\mathbf{S}(\mathbf{w}) \quad \to \quad \left(\sum_{t=1}^{T} w_t\right)\mathrm{E}(\mathbf{x}_t\mathbf{x}_t') = \mathbf{\Omega}_{xx}, \tag{27}$$

$$\mathbf{S}_1(\mathbf{w}_{(1)}) \quad \to \quad \left(\sum_{t=1}^{T_b} w_t\right)\mathrm{E}(\mathbf{x}_t\mathbf{x}_t') = \left(\sum_{t=1}^{T_b} w_t\right)\mathbf{\Omega}_{xx}, \tag{28}$$

and

$$\sum_{t=1}^{T} w_t^2\mathbf{x}_t\mathbf{x}_t' \to \left(\sum_{t=1}^{T} w_t^2\right)\mathbf{\Omega}_{xx}, \tag{29}$$

and the MSFE simplifies to

$$\mathrm{E}(\sigma_{(2)}^{-2}e_{T+1}^2) = 1 + (\mathbf{x}_{T+1}'\boldsymbol{\lambda})^2\left(\sum_{t=1}^{T_b} w_t\right)^2 + \mathbf{x}_{T+1}'\boldsymbol{\Omega}_{xx}^{-1}\mathbf{x}_{T+1}\left(\sum_{t=1}^{T_b} q^2 w_t^2 + \sum_{t=T_b+1}^{T} w_t^2\right). \quad (30)$$

The solution is similar to the case for $k = 1$ and is given by

$$w_{(1)} = \frac{1}{T}\frac{1}{b + (1-b)(q^2 + Tb\phi^2)}, \quad (31)$$

$$w_{(2)} = \frac{1}{T}\frac{q^2 + Tb\phi^2}{b + (1-b)(q^2 + Tb\phi^2)}, \quad (32)$$

where

$$\phi = \frac{\mathbf{x}_{T+1}'\boldsymbol{\lambda}}{\left(\mathbf{x}_{T+1}'\boldsymbol{\Omega}_{xx}^{-1}\mathbf{x}_{T+1}\right)^{1/2}}.$$

The above result is also in line with the result obtained for the simple case of $k = 1$. In that case $\boldsymbol{\Omega}_{xx} = \omega_x^2$ and $\phi = \lambda\omega_x$.

The above analysis assumes that the regressors are exogenous and, therefore, excludes dynamic models. The derivation of optimal weights for dynamic models is complicated by the fact that forecast errors are non-linear functions of past errors and, in general, it is difficult to obtain an analytic expression for the MSFE (Pesaran and Timmermann 2005). In order to see how the weights derived here perform in dynamic models, we have conducted Monte Carlo experiments, which show that using either optimal weights or the robust optimal weights developed in Section 3 leads to a substantial improvement in the forecast performance as compared to equal weights, post break sample, and AveW forecasts. To save space, MC results for dynamic models are provided in web supplement B.9.[7]

2.4 Multiple discrete breaks in a multiple regression model

Consider now the case of multiple breaks in the slope coefficients of a linear regression model

$$y_t = \boldsymbol{\beta}_t'\mathbf{x}_t + \sigma\varepsilon_t,$$

where the parameter vector $\boldsymbol{\beta}_t$ is subject to n breaks at points $b_i = T_{b,i}/T$, such that $b_1 < b_2 < \cdots < b_n$. For simplicity of exposition, we assume that the error variance is not subject to breaks. Initially, assume that $n = 2$, such that

$$\boldsymbol{\beta}_t = \begin{cases} \boldsymbol{\beta}_{(1)} & \text{for } 1 < t \leq T_{b,1} \\ \boldsymbol{\beta}_{(2)} & \text{for } T_{b,1} < t \leq T_{b,2} \\ \boldsymbol{\beta}_{(3)} & \text{for } T_{b,2} < t \leq T \end{cases}.$$

Using the weighted least squares estimator (19), we have that

$$\hat{\boldsymbol{\beta}}_T(\mathbf{w}) - \boldsymbol{\beta}_{(3)} = \mathbf{S}^{-1}(\mathbf{w})\left[\mathbf{S}_1(\mathbf{w}_{(1)})(\boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(3)}) + \mathbf{S}_2(\mathbf{w}_{(2)})(\boldsymbol{\beta}_{(2)} - \boldsymbol{\beta}_{(3)})\right] + \sigma\mathbf{S}^{-1}(\mathbf{w})\sum_{t=1}^{T} w_t\mathbf{x}_t\varepsilon_t,$$

[7] These MC results are also relevant to forecasts from regression models with serially correlated errors. This is because optimal forecasts from such models can be based on mathematically equivalent dynamic specifications with serially uncorrelated errors.

where $\mathbf{S}_1(\mathbf{w}_{(1)}) = \sum_{t=1}^{T_{b,1}} w_t \mathbf{x}_t \mathbf{x}_t'$, $\mathbf{S}_2(\mathbf{w}_{(2)}) = \sum_{t=T_{b,1}+1}^{T_{b,2}} w_t \mathbf{x}_t \mathbf{x}_t'$, and $\mathbf{S}(\mathbf{w}) = \sum_{t=1}^{T} w_t \mathbf{x}_t \mathbf{x}_t'$. Consequently,

$$\begin{aligned} e_{T+1}(\mathbf{w}) &= y_{T+1} - \mathbf{x}_{T+1}\hat{\boldsymbol{\beta}}_T(\mathbf{w}) \\ &= \sigma\varepsilon_{T+1} - \mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_1)(\boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(3)}) + \mathbf{S}_2(\mathbf{w}_{(2)})(\boldsymbol{\beta}_{(2)} - \boldsymbol{\beta}_{(3)})\right] \\ &\quad + \sigma\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \sum_{t=1}^{T} w_t \mathbf{x}_t \varepsilon_t, \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}\left[\sigma^{-2} e_{T+1}^2(\mathbf{w})\right] &= 1 + \left\{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}_{(1)} + \mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}_{(2)}\right]\right\}^2 \\ &\quad + \mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T} w^2 \mathbf{x}_t \mathbf{x}_t'\right) \mathbf{S}^{-1}(\mathbf{w}) \mathbf{x}_{T+1}, \end{aligned}$$

where $\boldsymbol{\lambda}_{(1)} = \left(\boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(3)}\right)/\sigma$ and $\boldsymbol{\lambda}_{(2)} = \left(\boldsymbol{\beta}_{(2)} - \boldsymbol{\beta}_{(3)}\right)/\sigma$.

In this case, the optimal weights can be obtained by solving the optimization problem

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{w}),$$

subject to $\boldsymbol{\iota}_T'\mathbf{w} = 1$, where

$$\begin{aligned} f(\mathbf{w}) &= \left\{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}_{(1)} + \mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}_{(2)}\right]\right\}^2 \\ &\quad + \mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T} w_t^2 \mathbf{x}_t \mathbf{x}_t'\right) \mathbf{S}^{-1}(\mathbf{w}) \mathbf{x}_{T+1}. \end{aligned} \tag{33}$$

The first order conditions are

$$\begin{aligned} &w_t \left[\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{T+1}\right] = \mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}_{(1)} + \mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}_{(2)}\right] \\ &\times \left\{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}_{(1)} + \mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}_{(2)}\right] - \mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\boldsymbol{\lambda}_{(i)}\right\} \\ &+ \mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T} w_t^2 \mathbf{x}_t \mathbf{x}_t'\right) \mathbf{S}^{-1}(\mathbf{w}) \mathbf{x}_{T+1} + \theta/2, \end{aligned}$$

where again θ is the Lagrange multiplier associated with $\boldsymbol{\iota}_T'\mathbf{w} = 1$ and

$$\boldsymbol{\lambda}_{(i)} = \begin{cases} \boldsymbol{\lambda}_{(1)} & \text{if } t \leq T_{b,1} \\ \boldsymbol{\lambda}_{(2)} & \text{if } T_{b,1} < t \leq T_{b,2} \\ \mathbf{0} & \text{if } T_{b,2} < t \leq T \end{cases}.$$

Again, by multiplying both sides by w_t and summing over $t = 1, 2, \ldots, T$ it can be easily verified that $\theta = 0$. Hence, for $\mathbf{x}_t \neq \mathbf{0}$ the optimal weights are

$$\begin{aligned} w_t &= \frac{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}_{(1)} + \mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}_{(2)}\right]}{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t} \times \\ &\quad \times \left\{\mathbf{x}_t'\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}_{(1)} + \mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}_{(2)}\right]\right\} \\ &\quad + \frac{\mathbf{x}_t'\mathbf{S}^{-1}(\mathbf{w}) \left(\sum_{t=1}^{T} w_t^2 \mathbf{x}_t \mathbf{x}_t'\right) \mathbf{S}^{-1}(\mathbf{w}) \mathbf{x}_{t+1}}{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t} \\ &\quad - \frac{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w}) \left[\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}_{(1)} + \mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}_{(2)}\right] (\mathbf{x}_t'\boldsymbol{\lambda}_{(i)})}{\mathbf{x}_{T+1}'\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t}. \end{aligned} \tag{34}$$

For n breaks we obtain

$$
\begin{aligned}
w_t &= \frac{\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\left[\sum_{j=1}^{n-1}\mathbf{S}_j(\mathbf{w}_{(j)})\boldsymbol{\lambda}_{(j)}\right]\left\{\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\left[\sum_{j=1}^{n-1}\mathbf{S}_j(\mathbf{w}_{(j)})\boldsymbol{\lambda}_{(j)}\right]\right\}}{\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t} \\
&+\frac{\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\left(\sum_{t=1}^{T}w_t^2\mathbf{x}_t\mathbf{x}'_t\right)\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{t+1}}{\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t}-\frac{\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\left[\sum_{j=1}^{n-1}\mathbf{S}_j(\mathbf{w}_{(j)})\boldsymbol{\lambda}_{(j)}\right](\mathbf{x}'_t\boldsymbol{\lambda}_{(i)})}{\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_t},
\end{aligned}
$$

where $\boldsymbol{\lambda}_{(i)}=\left(\boldsymbol{\beta}_{(i)}-\boldsymbol{\beta}_{(n+1)}\right)/\sigma$ if $T_{b,i-1}<t\leq T_{b,i}$, with $T_{b,0}=1$, and $\boldsymbol{\lambda}_{(n+1)}=\mathbf{0}$. As in the case of a single break, numerical methods are necessary to obtain the weights but, again, in the case of a single regressor or asymptotically we can derive analytical results.

2.4.1 Optimal weights for multiple breaks in a simple regression model

In the case of a single regressor, (34) simplifies to

$$
w_t=\frac{[S_1(\mathbf{w}_{(1)})\lambda_{(1)}+S_2(\mathbf{w}_{(2)})\lambda_{(2)}]^2}{S(\mathbf{w})}+\frac{\sum_{t=1}^{T}w_t^2x_t^2}{S(\mathbf{w})}-[S_1(\mathbf{w}_{(1)})\lambda_{(1)}+S_2(\mathbf{w}_{(2)})\lambda_{(2)}]\lambda_{(i)},
$$

where $\lambda_{(i)}$ is as defined above but it is now a scalar. Therefore, defining $S_1(1)=\sum_{t=1}^{T_{b,1}}x_t^2$ and $S_2(1)=\sum_{t=T_{b,1}+1}^{T_{b,2}}x_t^2$, solving for the optimal weights yields

$$
\begin{aligned}
w_{(1)} &= \frac{1}{T}\frac{1+\lambda_{(2)}^2S_2(1)-\lambda_{(1)}\lambda_{(2)}S_2(1)}{a_{s,2}}, \\
w_{(2)} &= \frac{1}{T}\frac{1+\lambda_{(1)}^2S_1(1)-\lambda_{(1)}\lambda_{(2)}S_1(1)}{a_{s,2}}, \\
w_{(3)} &= \frac{1}{T}\frac{1+\lambda_{(1)}^2S_1(1)+\lambda_{(2)}^2S_2(1)}{a_{s,2}},
\end{aligned}
$$

where $a_{s,2}=1+(1-b_2)\left[S_1(1)\lambda_{(1)}^2+S_2(1)\lambda_{(2)}^2\right]+\left[\lambda_{(1)}-\lambda_{(2)}\right]\left[(b_2-b_1)S_1(1)\lambda_{(1)}-b_1S_2(1)\lambda_{(2)}\right]$. This result generalizes to n breaks where

$$
\begin{aligned}
w_{(i)}|i\leq n &= \frac{1}{T}\frac{1+\sum_{j=1,j\neq i}^{n}\lambda_{(j)}^2S_j(1)-\lambda_{(i)}\sum_{j=1,j\neq i}^{n}\lambda_{(j)}S_j(1)}{a_{s,n}}, \\
w_{(n+1)} &= \frac{1}{T}\frac{1+\sum_{j=1}^{n}\lambda_{(j)}^2S_j(1)}{a_{s,n}},
\end{aligned}
$$

and $a_{s,n}=1+\sum_{l=1}^{n+1}(b_l-b_{l-1})\sum_{j=1,j\neq l}^{n}\lambda_{(j)}^2S_j(1)-\sum_{l=1}^{n}\lambda_{(l)}(b_l-b_{l-1})\sum_{j=1,j\neq l}^{n}\lambda_{(j)}S_j(1)$.

2.4.2 Asymptotic results in the multi-break case with $k\geq 1$ stationary regressors

Similar to the case with one break, we can simplify the solution when there are two or more regressors if we assume that many observations are available between breaks, and $\mathbf{x}_t$ is a stationary process with $\mathrm{E}(\mathbf{x}_t\mathbf{x}'_t)=\boldsymbol{\Omega}_{xx}$. Note, however, that we make no

assumption about the number of observations since the last break. Initially consider the case of two breaks. In addition to (27) and (29) we have

$$\mathbf{S}_1(\mathbf{w}_{(1)}) \to \left(\sum_{t=1}^{T_{b,1}} w_t\right) \mathrm{E}(\mathbf{x}_t\mathbf{x}_t') = \left(\sum_{t=1}^{T_{b,1}} w_t\right) \mathbf{\Omega}_{xx},$$
$$\mathbf{S}_2(\mathbf{w}_{(2)}) \to \left(\sum_{t=T_{b,1}+1}^{T_{b,2}} w_t\right) \mathrm{E}(\mathbf{x}_t\mathbf{x}_t') = \left(\sum_{t=T_{b,1}+1}^{T_{b,2}} w_t\right) \mathbf{\Omega}_{xx}.$$

Then (33) simplifies to

$$f(\mathbf{w}) = \left[\mathbf{x}_{T+1}' \left(\boldsymbol{\lambda}_{(1)} \sum_{t=1}^{T_{b,1}} w_t + \boldsymbol{\lambda}_{(2)} \sum_{t=T_{b,1}+1}^{T_{b,2}} w_t\right)\right]^2 + \mathbf{x}_{T+1}' \sum_{t=1}^{T} w_t^2 \mathbf{\Omega}_{xx}^{-1} \mathbf{x}_{T+1}.$$

The optimal weights are therefore

$$w_{(1)} = \frac{1}{T} \frac{1 + T(b_2 - b_1)\phi_{(2)}^2 - T(b_2 - b_1)\phi_{(1)}\phi_{(2)}}{a_{a,2}}, \tag{35}$$
$$w_{(2)} = \frac{1}{T} \frac{1 + Tb_1\phi_{(1)}^2 - Tb_1\phi_{(1)}\phi_{(2)}}{a_{a,2}}, \tag{36}$$
$$w_{(3)} = \frac{1}{T} \frac{1 + Tb_1\phi_{(1)}^2 + T(b_2 - b_1)\phi_{(2)}^2}{a_{a,2}}, \tag{37}$$

where $a_{a,2} = 1 + T(1 - b_2)b_1\phi_{(1)}^2 + T(b_2 - b_1)(1 - b_2)\phi_{(2)}^2 + Tb_1(b_2 - b_1)(\phi_{(1)} - \phi_{(2)})^2$ and

$$\phi_{(i)} = \frac{\mathbf{x}_{T+1}'\boldsymbol{\lambda}_{(i)}}{\left(\mathbf{x}_{T+1}'\mathbf{\Omega}_{xx}^{-1}\mathbf{x}_{T+1}\right)^{1/2}}, \text{ for } i = 1, 2.$$

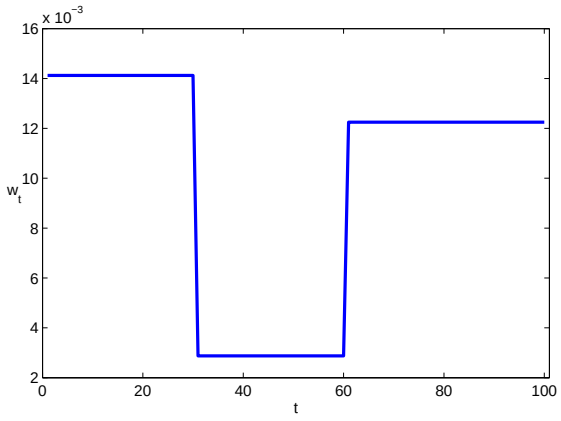
An interesting result is that the weights for two breaks are not necessarily decreasing in the distance from the end point, T. In particular,

- $w_{(1)} > w_{(3)} > w_{(2)}$ if $\phi_{(1)} < 0$, $\phi_{(2)} > 0$ and $b_1\phi_{(1)} > -(b_2 - b_1)\phi_{(2)}$,
- $w_{(1)} > w_{(3)} > w_{(2)}$ if $\phi_{(1)} > 0$, $\phi_{(2)} < 0$ and $b_1\phi_{(1)} < -(b_2 - b_1)\phi_{(2)}$,
- $w_{(2)} > w_{(3)} > w_{(1)}$ if $\phi_{(1)} < 0$, $\phi_{(2)} > 0$ and $b_1\phi_{(1)} < -(b_2 - b_1)\phi_{(2)}$,
- $w_{(2)} > w_{(3)} > w_{(1)}$ if $\phi_{(1)} > 0$, $\phi_{(2)} < 0$ and $b_1\phi_{(1)} > -(b_2 - b_1)\phi_{(2)}$.

Figure 1 plots the weights for $T = 100$, $b_1 = 0.3$, $b_2 = 0.6$, $\phi_{(1)} = -0.5$ and $\phi_{(2)} = 1.5$. Under this parameter constellation it is easily seen that $w_{(1)} > w_{(3)} > w_{(2)}$. This result may be surprising at first sight. The intuition is that the observations after the last break deliver an unbiased forecast. In contrast, the observations before the last break will generally introduce a bias into the forecast. Biases of opposite sign can offset each other but if one bias is larger in absolute terms the observations in the remaining sub-sample must receive a larger weight to offset this bias such that the MSFE is minimized.

Note that the weights $w_{(1)}$ and $w_{(2)}$ can be negative. We do not restrict the weights to be positive as the weights in (35), (36), and (37) give a unique minimum of the MSFE.

Figure 1: Optimal weights for $T = 100$, $b_1 = 0.3$, $b_2 = 0.6$, $\phi_{(1)} = -0.5$ and $\phi_{(2)} = 1.5$



In the case of n breaks, the weights for the $n+1$ segments are given by

$$w_{(i)|i\leq n} = \frac{1}{T}\frac{1 + T\sum_{j=1,j\neq i}^{n}(b_j - b_{j-1})\phi_{(j)}^2 - T\phi_{(i)}\sum_{j=1,j\neq i}^{n}(b_j - b_{j-1})}{a_{a,n}}, \tag{38}$$

$$w_{(n+1)} = \frac{1}{T}\frac{1 + T\sum_{j=1}^{n}(b_j - b_{j-1})\phi_{(j)}^2}{a_{a,n}}, \tag{39}$$

where $a_{a,n} = 1+T\sum_{l=1}^{n+1}(b_l-b_{l-1})\sum_{j=1,j\neq l}^{n}\phi_{(j)}^2(b_j-b_{j-1})-T\sum_{l=1}^{n}\phi_{(l)}(b_l-b_{l-1})\sum_{j=1,j\neq l}^{n}\phi_{(j)}^2(b_j-b_{j-1})$ and $b_0 = 0$. Expressions of the weights in matrix notation that are convenient when programming the weights can be found in the web supplement B.4.

3 Optimal weights when the time and size of the break are uncertain

So far we have assumed that the time and the size of the break are known. However, this may not be the case in many situations of practical interest. In particular, the size of the break is difficult to estimate unless a relatively large number of post-break observations is available.[8] It is, therefore, worthwhile to develop weights that are reasonably robust to the point and the size of the break(s). As a simple example, consider the model with a single break at time T_b both in the slopes and the error variances. Using (24) and (25) we first note that

$$Tw_{(1)} = \frac{1}{b + (1-b)q^2 + Tb(1-b)\phi^2}, \quad \text{and} \quad Tw_{(2)} = \frac{q^2 + Tb\phi^2}{b + (1-b)q^2 + Tb(1-b)\phi^2},$$

where $\phi^2 = \lambda^2\hat{\omega}_x^2$, with $\hat{\omega}_x^2 = T^{-1}\sum_{t=1}^{T}x_t^2$, $\lambda = (\beta_{(1)}-\beta_{(2)})/\sigma_{(2)}$ and $q = \sigma_{(1)}/\sigma_{(2)}$. The time profile of the weights can be written as $Tw_t(b, q^2, \phi^2) = w_{(2)} + \left[w_{(1)} - w_{(2)}\right]\mathrm{I}(T_b - t)$, for $t = 1, 2, \ldots, T$. Hence

$$Tw(a, b, q^2, \phi^2) = \frac{\frac{q^2}{T} + b\phi^2}{\frac{b+(1-b)q^2}{T} + b(1-b)\phi^2} + \left[\frac{\frac{1-q^2}{T} - b\phi^2}{\frac{b+(1-b)q^2}{T} + b(1-b)\phi^2}\right]\mathrm{I}(b - a), \tag{40}$$

[8] Also in finite samples the distribution of the estimated break point does not have a closed form expression and depends on the distribution of x_t and ε_t. (See Hinkley, 1970). Asymptotic results can be obtained that do not depend on the distribution of the regressors or the error term (e.g. Bai 1997), but such results might not be reliable in small samples.

where $a = t/T \in [0,1]$, and as before $b = T_b/T \in [0,\bar{b}]$, where $\bar{b} < 1$.

Initially, consider the case where the break is in the error variances only, namely $\phi = 0$ and $q^2 \neq 1$. Then

$$Tw(a,b,q^2) = \frac{q^2}{b+(1-b)q^2} + \left[\frac{1-q^2}{b+(1-b)q^2}\right] \mathrm{I}(b-a),$$

or

$$Tw(a,b,q^2) = \frac{1}{1+b\psi} + \left(\frac{\psi}{1+b\psi}\right) \mathrm{I}(b-a).$$

where $\psi = (1-q^2)/q^2 = (\sigma^2_{(2)} - \sigma^2_{(1)})/\sigma^2_{(1)}$. It is also worth noting that $w_{(1)}/w_{(2)} = 1+\psi = \sigma^2_{(2)}/\sigma^2_{(1)}$, and more weight will be given to pre break observations if $\sigma^2_{(2)} > \sigma^2_{(1)}$, and *vice versa*. This is in line with the result obtained by Pesaran and Timmermann (2007) using the concept of the optimal window.

In situations where b and q^2 are uncertain their effects on the optimal weights can be integrated out with respect to a given distribution of b and q^2. Here, we assume that b and q^2 are independently distributed and focus on the uncertainty of b for a given value of q^2, or ψ. For b we assume that it is uniformly distributed over the range $\underline{b}$ and $\bar{b}$, namely the probability density of b is given by

$$f(b) = \begin{cases} 0 & \text{if } b < \underline{b} \\ (\bar{b}-\underline{b})^{-1} & \text{if } \underline{b} \leq b < \bar{b} \\ 0 & \text{if } b \geq \bar{b}. \end{cases}.$$

The expression for $w(a,q^2)$ depends on whether a falls within the range $[\underline{b},\bar{b}]$ or not. Specifically, we have

$$Tw(a,q^2) = \begin{cases} (\bar{b}-\underline{b})^{-1} \int_{\underline{b}}^{\bar{b}} \frac{1+\psi}{1+b\psi} db & \text{if } a < \underline{b} \\ (\bar{b}-\underline{b})^{-1} \int_{\underline{b}}^{\bar{b}} \frac{1}{1+b\psi} db + \frac{\psi}{\bar{b}-\underline{b}} \int_{a}^{\bar{b}} \frac{1}{1+b\psi} db & \text{if } \underline{b} \leq a \leq \bar{b} \\ (\bar{b}-\underline{b})^{-1} \int_{\underline{b}}^{\bar{b}} \frac{1}{1+b\psi} db & \text{if } a > \bar{b} \end{cases}.$$

Also, it is easily seen that

$$\int_{\underline{b}}^{\bar{b}} \frac{1}{1+b\psi} db = \psi^{-1} \log\left(\frac{1+\bar{b}\psi}{1+\underline{b}\psi}\right),$$

and, hence,

$$Tw(a,q^2) = (\bar{b}-\underline{b})^{-1} \left[\psi^{-1} \log\left(\frac{1+\bar{b}\psi}{1+\underline{b}\psi}\right) + \log\left(\frac{1+\bar{b}\psi}{1+a\psi}\right)\right], \quad \text{if } \underline{b} \leq a \leq \bar{b}.$$

Since $\psi = (1-q^2)/q^2$, we can also write

$$Tw(a,q^2) = \begin{cases} (\bar{b}-\underline{b})^{-1} \frac{1}{1-q^2} \log\left(\frac{\bar{b}+(1-\bar{b})q^2}{\underline{b}+(1-\underline{b})q^2}\right) & \text{if } a < \underline{b} \\ (\bar{b}-\underline{b})^{-1} \left[\frac{q^2}{1-q^2} \log\left(\frac{\bar{b}+(1-\bar{b})q^2}{\underline{b}+(1-\underline{b})q^2}\right) + \log\left(\frac{\bar{b}+(1-\bar{b})q^2}{a+(1-a)q^2}\right)\right] & \text{if } \underline{b} \leq a \leq \bar{b} \\ (\bar{b}-\underline{b})^{-1} \frac{q^2}{1-q^2} \log\left(\frac{\bar{b}+(1-\bar{b})q^2}{\underline{b}+(1-\underline{b})q^2}\right) & \text{if } a > \bar{b} \end{cases} \tag{41}$$

Over the range $\underline{b} \leq a \leq \bar{b}$

$$\frac{T\partial w(a,q^2)}{\partial a} = (\bar{b}-\underline{b})^{-1} \frac{-(1-q^2)}{a+(1-a)q^2} = \frac{1}{\bar{b}-\underline{b}} \frac{\sigma^2_{(1)} - \sigma^2_{(2)}}{\sigma^2_{(2)}\left[a+(1-a)q^2\right]}.$$

and the weights $w(a, q^2)$ monotonically rise (fall) with a if $\sigma^2_{(1)} > \sigma^2_{(2)}$ ($\sigma^2_{(1)} < \sigma^2_{(2)}$). In other words, more weight will be placed on more recent observations only if post-break error variance is smaller than pre-break error variance. This result holds for all values of T.

Consider now the more general case where $\phi^2 > 0$. Using (40) we have for $\underline{b} < a < \bar{b}$

$$Tw(a, b, q^2, \phi^2) = \frac{\frac{q^2}{\phi^2 T} + b}{\frac{b+(1-b)q^2}{T\phi^2} + b(1-b)} + \frac{\frac{1-q^2}{\phi^2 T} - b}{\frac{b+(1-b)q^2}{T\phi^2} + b(1-b)} \mathrm{I}(b-a),$$

For given values of q^2 and ϕ^2 and assuming that b lies in the range $[\underline{b}, \bar{b}]$ with $0 < \underline{b} < \bar{b} < 1$, we have for $\underline{b} < a < \bar{b}$

$$Tw(a \,|q^2, \phi^2) = \int_{\underline{b}}^{\bar{b}} \frac{\frac{q^2}{\phi^2 T} + b}{\frac{[q^2+(1-q^2)b]}{T\phi^2} + b(1-b)} db + \int_{a}^{\bar{b}} \frac{\frac{1-q^2}{\phi^2 T} - b}{\frac{[b+(1-b)q^2]}{T\phi^2} + b(1-b)} db, \qquad (42)$$

It is now easily seen that

$$T\frac{\partial w(a \,|\phi^2, q^2)}{\partial a} = \frac{-(\frac{1-q^2}{\phi^2 T} - a)}{\frac{[a+(1-a)q^2]}{T\phi^2} + a(1-a)}.$$

that is, the weights increase monotonically in a if $a > \frac{1-q^2}{\phi^2 T}$, which is clearly satisfied if $q^2 \geq 1$. In this case, the observations farthest from the end of the sample get the smallest weights. The decay rate of the weights depends on T.

3.1 Large T approximation

Consider now a large T approximation of the optimal weights and note that

$$Tw(a, b, q^2, \phi^2) = \frac{\frac{q^2}{\phi^2 T} + b}{b(1-b)\left(1 + \frac{\theta}{T}\right)} + \frac{\left(\frac{1-q^2}{\phi^2 T} - b\right) \mathrm{I}(b-a)}{b(1-b)\left(1 + \frac{\theta}{T}\right)},$$

where $\theta = \left[q^2 + (1-q^2)b\right] / \phi^2 b(1-b) > 0$. Using $\left(1 + \frac{\theta}{T}\right)^{-1} = 1 - \frac{\theta}{T} + O(T^{-2})$, and replacing θ in terms of b, q, and ϕ, we have

$$\begin{aligned} Tw(a, b, q^2, \phi^2) &= \frac{1}{1-b} - \frac{1}{1-b}\mathrm{I}(b-a) + \frac{1}{T}\left[\frac{q^2}{\phi^2 b(1-b)} - \frac{q^2 + (1-q^2)b}{\phi^2 b(1-b)^2}\right] \\ &\quad + \frac{1}{T}\left[\frac{1-q^2}{\phi^2 b(1-b)} + \frac{q^2 + (1-q^2)b}{\phi^2 b(1-b)^2}\right] \mathrm{I}(b-a) + O(T^{-2}). \end{aligned}$$

But since

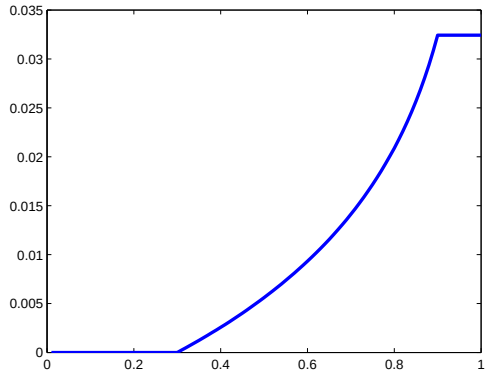
$$\frac{q^2}{\phi^2 b(1-b)} - \frac{q^2 + (1-q^2)b}{\phi^2 b(1-b)^2} = \frac{q^2(1-b) - q^2 - (1-q^2)b}{\phi^2 b(1-b)^2} = \frac{-1}{\phi^2(1-b)^2},$$

$$\frac{1-q^2}{\phi^2 b(1-b)} + \frac{q^2 + (1-q^2)b}{\phi^2 b(1-b)^2} = \frac{\left(1-q^2\right)(1-b) + q^2 + (1-q^2)b}{\phi^2 b(1-b)^2} = \frac{1}{\phi^2 b(1-b)^2},$$

the weights profile simplifies to

$$\begin{aligned} Tw(a, b, q^2, \phi^2) &= \frac{1}{1-b} - \frac{1}{1-b}\mathrm{I}(b-a) \qquad (43) \\ &\quad - \frac{1}{T}\left[\frac{1}{\phi^2(1-b)^2}\right] + \frac{1}{T}\left[\frac{1}{\phi^2 b(1-b)^2}\right] \mathrm{I}(b-a) + O(T^{-2}). \end{aligned}$$

Figure 2: Robust optimal weights (44), $T = 100, \underline{b} = 0.3, \bar{b} = 0.9$



It is interesting that the first order term in this expansion does not depend on the sizes of the breaks, and depends only on the break point, b. Also, the terms up to order T^{-1} are independent of q^2 as long as $\phi^2 > 0$, that is, a break in the error variance is dominated by a break in the mean of the process.

Therefore, for large T, robust optimal weights are determined by the distribution of b. For the uniform distribution, $b \sim \text{Uniform}(\underline{b}, \bar{b})$ with $0 < \underline{b} < \bar{b} < 1$, we have

$$Tw(a) = \begin{cases} 0 + O(T^{-1}), & \text{for } a < \underline{b} \\ (\bar{b} - \underline{b})^{-1} \int_{\underline{b}}^{\bar{b}} \frac{1}{1-b} db - (\bar{b} - \underline{b})^{-1} \int_{a}^{\bar{b}} \frac{1}{1-b} db + O(T^{-1}), & \text{for } \underline{b} \leq a \leq \bar{b} \\ (\bar{b} - \underline{b})^{-1} \int_{\underline{b}}^{\bar{b}} \frac{1}{1-b} db + O(T^{-1}), & \text{for } a > \bar{b} \end{cases},$$

and the robust optimal weights are

$$w(a) \approx \begin{cases} 0, & \text{if } a < \underline{b} \\ \frac{-1}{T(\bar{b}-\underline{b})} \log\left(\frac{1-a}{1-\underline{b}}\right), & \text{if } \underline{b} \leq a \leq \bar{b} \\ \frac{-1}{T(\bar{b}-\underline{b})} \log\left(\frac{1-\bar{b}}{1-\underline{b}}\right), & \text{if } a > \bar{b} \end{cases}. \tag{44}$$

Figure 2 shows the robust optimal weights for $T = 100$, $\underline{b} = 0.3$ and $\bar{b} = 0.9$, assuming that $\phi^2 > 0$, and it can be seen that the weights increase monotonically from $\underline{b}$ to $\bar{b}$.

In the case where $\underline{b}$ and $\bar{b}$ are close to the end points of 0 and 1, we have

$$w(a) \approx \frac{-\log(1-a)}{T}, \; a \in \left[0, \bar{b}\right]. \tag{45}$$

A discrete time version can be obtained by setting $T\bar{b} = T - 1$, or $\bar{b} = 1 - 1/T$.[9] This gives robust optimal weights that integrate the break date over the entire estimation sample

$$w_t^* = \frac{-\log(1 - t/T)}{T - 1}, \text{ for } t = 1, 2, \ldots, T-1, \tag{46}$$

$$w_T^* = \frac{-1}{T-1} \log\left(1 - \frac{T-1}{T}\right) = \frac{\log(T)}{T-1}. \tag{47}$$

[9]Clearly, one could set $\bar{b}$ to other values close to 1, say $1 - 0.5/T$. But for relatively large T, the choice of w_T^* for the forecasts is unlikely to be of great importance.

Due to the approximation/discretization these weights do not sum to unity, and can be scaled as

$$w_t = \frac{w_t^*}{\sum_{s=1}^{T} w_s^*}, \text{ for } t = 1, 2, \ldots, T. \tag{48}$$

For $\underline{b}$ and $\bar{b}$ close to the end points of 0 and 1, we can also obtain the MSFE implied by the robust optimal weights (44) as[10]

$$\begin{aligned}\frac{\omega_x^2}{x_{T+1}^2}\left[\mathrm{E}\left(\sigma_{(2)}^{-2}e_{T+1}^2\right)-1\right] &\approx \phi^2\left[b+(1-b)\log(1-b)\right]^2 \\ &+\frac{\left(q^2-1\right)}{T}\left[-(1-b)\left[\log(1-b)\right]^2+2(1-b)\log(1-b)+2b\right]+\frac{2}{T}.\end{aligned} \tag{49}$$

Comparing this result to the equal weight MSFE we have

$$\frac{\omega_x^2}{x_{T+1}^2}\left(\mathrm{MSFE}_{equal}-\mathrm{MSFE}_{robust}\right)=\phi^2\psi_\phi(b)+\frac{\left(q^2-1\right)}{T}\psi_q(b)-\frac{1}{T},$$

where

$$\psi_\phi(b)=\left[b^2-\left[b+(1-b)\log(1-b)\right]^2\right]=\left[2b+(1-b)\log(1-b)\right]\left[-(1-b)\log(1-b)\right],$$

and $\psi_q(b) = (1-b)\left[\log(1-b)\right]^2 - 2(1-b)\log(1-b) - b$. The relative performance of the two sets of weights depend on the sign of $\left(q^2-1\right)\psi_q(b)$. It can be shown that $\psi_q(b) > 0$ if $b \leq 0.91$, and negative otherwise. However, for reasonable values of q^2 (say 1/2 or 2), the term $\frac{\left(q^2-1\right)}{T}\psi_q(b)$ is relatively unimportant when T is 100 or more. Note that $\max_{0\leq b\leq 0.95}|\psi_q(b)| = 0.202$ and for $T = 100$ the contribution of $\frac{\left(q^2-1\right)}{T}\psi_q(b)$ to the relative performance of the two weights can be ignored, unless ϕ is very small and b very close to 0 or 1.

It is also interesting to compare the fit of the robust optimal weights and the ExpS weights to the optimal weights for a range of T, q^2 and ϕ^2. Figure 3 contains plots of the optimal weights $w(a|q^2,\phi^2)$ in (42), the robust optimal weights, w_t^*, in (48) and the ExpS weights (10), where γ is chosen such that the distance between $w(a|q^2,\phi^2)$ and $w_t^e(\gamma)$ is minimized. The plots show that the accuracy of the robust optimal weights depends largely on ϕ^2: for larger ϕ^2 the robust optimal weights are very close to the optimal weights, for the smaller ϕ^2 a good approximation requires large T. The plots also show that, as predicted by our theory, q^2 has a relatively minor influence on the weights that is only visible when T and ϕ^2 are both small, which is visible in the top left plot, where initially the weights fall slightly. Finally, the down-weighting parameter γ in the exponential smoothing weight that best approximates the exact optimal weight varies between 0.944 and 0.994, and the ExpS weights generally give too low a weight to the most recent observations as compared to the optimal weights.[11]

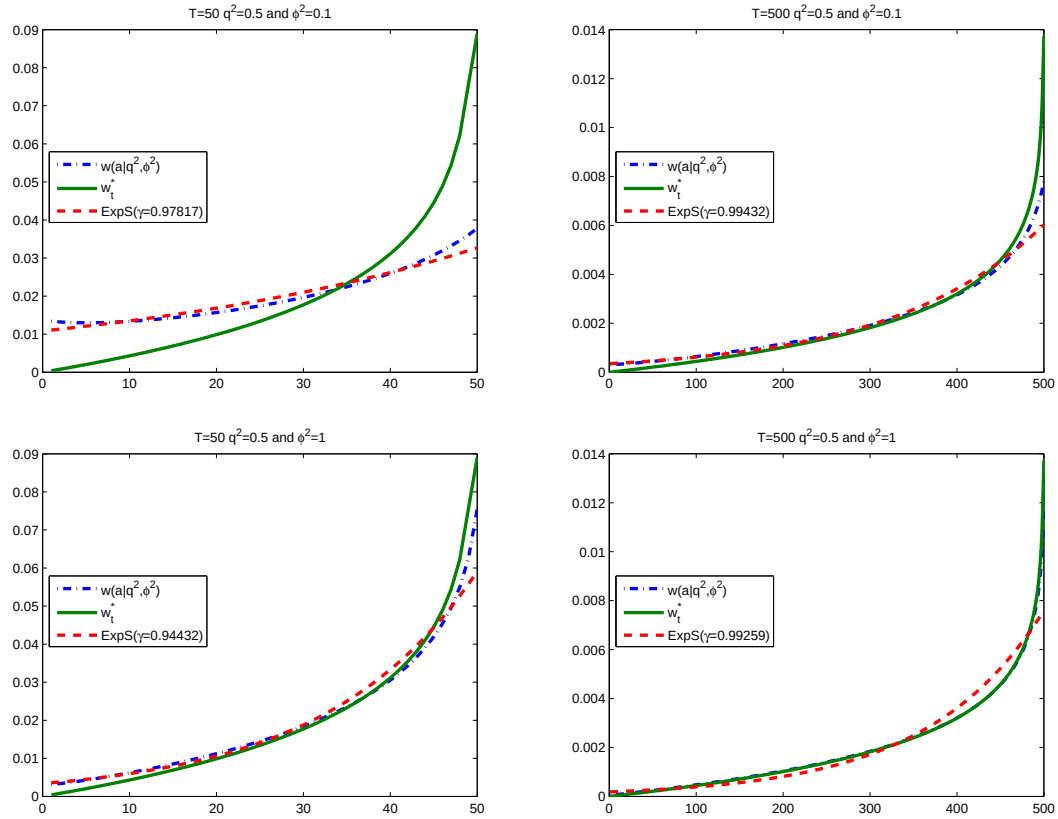
3.1.1 Robust optimal weights for regression models with two breaks

Consider the case of two breaks, where the weights conditional on b and λ are given in (35) to (37). Clearly, $\underline{b} < b_1 < b_2 < \bar{b}$ and $\Pr(b_1, b_2) = \Pr(b_1)\Pr(b_2|b_1)$, furthermore

[10] Derivations for the MSFE can be found in web supplement B.5.

[11] It is possible to derive robust optimal weights that allow for high order terms in the expansion (43), and some results are provided in web supplement B.7. However, we will not pursue them further in this paper.

Figure 3: Comparison of optimal weights, robust optimal weights, and fitted exponential smoothing weights



Note: $T = 50$ in the plots in the left column, $T = 500$ in the plots in the right column, $\phi^2 = 0.1$ in the plots in the top row, $\phi^2 = 1$ in the plots in the bottom row, and $q^2 = 0.5$ throughout. The dash-dotted line represents the optimal weights $w(a|q^2, \phi^2)$ in (42), the solid line the robust optimal weights in (48), and the dashed line the ExpS weights in (10).

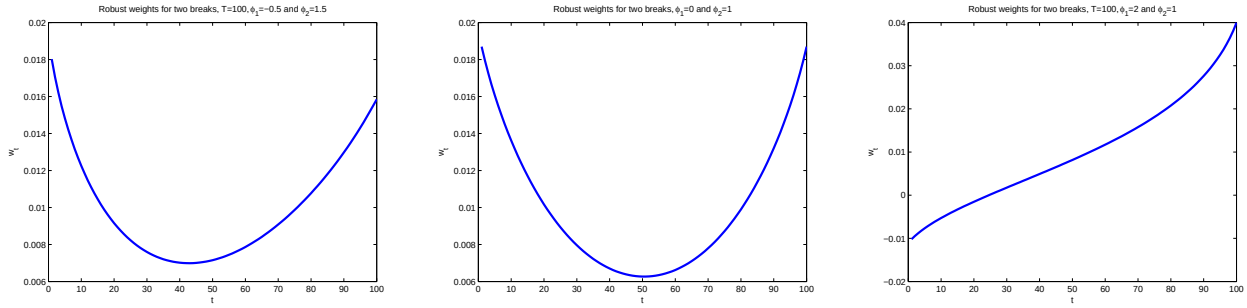
$\underline{b}_1 < b_1 < \bar{b}_1$ and $\underline{b}_2 < b_2 < \bar{b}_2$ where $\underline{b}_1 < \underline{b}_2$ and $\bar{b}_1 < \bar{b}_2$, then

$$\Pr(b_1) = \begin{cases} 0 & \text{if } b_1 < \underline{b}_1 \\ \frac{1}{\bar{b}_1 - \underline{b}_1} & \text{if } \underline{b}_1 < b_1 \leq \bar{b}_1 \\ 0 & \text{if } b_1 > \bar{b}_1 \end{cases}, \quad \text{and} \quad \Pr(b_2|b_1) = \begin{cases} 0 & \text{if } b_2 < \underline{b}_2 \\ \frac{1}{\bar{b}_2 - b_1} & \text{if } \underline{b}_2 < b_2 \leq \bar{b}_2 \\ 0 & \text{if } b_2 > \bar{b}_2 \end{cases}.$$

Analytic solutions for the robust optimal weights under two breaks are not easy to obtain. However, the weights can be calculated numerically using (35) to (37) and integrating over a grid for b_1 and b_2 taking into account that $b_1 < b_2$ and setting $\underline{b}_1 = 1/T$, $\underline{b}_2 = 2/T$, $\bar{b}_1 = (T-2)/T$ and, finally, $\bar{b}_2 = (T-1)/T$.

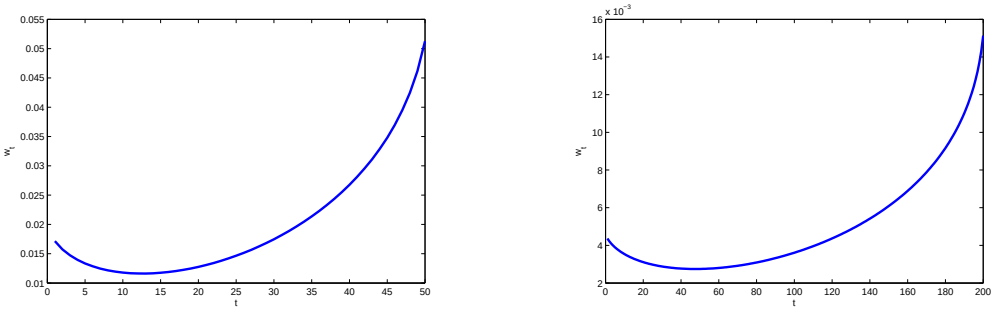
Figure 4 plots the robust optimal weights for two breaks and $T = 100$, where the first graph reports the weights for $\phi_{(1)} = -0.5$ and $\phi_{(2)} = 1.5$, the second for $\phi_{(1)} = 0$ and $\phi_{(2)} = 1$, the third for $\phi_{(1)} = 2$ and $\phi_{(2)} = 1$. It can be seen that the shape of the weights depends on the parameters chosen. In the first graph, the parameters $\phi_{(1)}$ and $\phi_{(2)}$ are those that under known break dates resulted in the example in Figure 1 where the first sub-sample receives the largest weights. The pattern is the same with the very early observation receiving higher weights than the last observations. The second graph is for parameters that would lead to equal weights in the first and last

Figure 4: Robust optimal weights for two breaks, $T = 100$, and different values of $\phi_{(1)}$ and $\phi_{(2)}$



Note: The first graph plots the weights for $\phi_{(1)} = -0.5$ and $\phi_{(2)} = 1.5$, the second for $\phi_{(1)} = 0$ and $\phi_{(2)} = 1$, and the third for $\phi_{(1)} = 2$ and $\phi_{(2)} = 1$. The weights are given in (35) to (37) and integrating uniformly over b_1 and b_2 over the range $1/T$ to $(T-1)/T$.

Figure 5: Robust optimal weights for two breaks and $\phi_{(1)}$ and $\phi_{(2)}$ integrated out, $T = 50, 200$



Note: The first graph plots the weights for $T = 50$ and the second for $T = 200$. The weights are given in (35) to (37) and integrating uniformly over b_1 and b_2 over the range $1/T$ to $(T-1)/T$ and $\phi_{(1)}$ and $\phi_{(2)}$ over the range -2 to 2.

sub-sample if the break dates were known. The final graph uses breaks that decrease in size, which results in continuously increasing weights.

In practice, given that the break date is uncertain, the size of break is also likely to be unknown. In addition to the break date, we therefore also integrate over the break sizes. Figure 5 plots the weights when $\phi_{(1)}$ and $\phi_{(2)}$ are integrated with respect to a uniform distribution in the range -2 to 2. The first graph shows the weights for $T = 50$ and the second for $T = 200$. It can be seen that the shape of the weight function is largely independent of the sample size. Most weight is given to the most recent observations. Interestingly, the first observations receive larger weights than the observations in the middle of the sample, which reflects the possibility that early observations can have a bias that counterbalances that of later observations.

4 Monte Carlo evidence on forecasting performance

4.1 Data generating process

We now provide a range of Monte Carlo experiments, comparing the forecast performance of the optimal weights and robust optimal weights proposed in this paper as compared to other alternatives available in the literature. The first set of experiments considers the continuous break model (2) in Section 2.1. A second set of experiments concentrates on the random walk model (2) with a single discrete break as discussed in Section 2.2. In this model, the MSFEs of the different forecasting methods are known conditional on T_b and λ and have been reported in Table 1. In the case of these experiments we aim to find out how far the uncertainty around the break date and size affects the different forecasts. In a final set of experiments we add a regressor using the simple linear regression model discussed in Section 2.3.1.

The model for the first two experiments is

$$y_t = \mu_t + \sigma_\varepsilon \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0,1). \tag{50}$$

In the first set of experiments, the mean follows the random walk specification

$$\mu_t = \mu_{t-1} + \sigma_v v_t, \quad v_t \sim \mathrm{N}(0,1),$$

for $t = 1, 2, \ldots, T, T+1$ with $T = 50, 100, 200$, and $\gamma = \{0.8, 0.9, 0.95, 0.98\}$, which corresponds to $\delta = \sigma_\varepsilon/\sigma_v \approx \{4.472, 9.487, 19.494, 49.497\}$. For the second set of experiments, the mean in (50) has a discrete break

$$\mu_t = \begin{cases} \mu_{(1)} & t \leq T_b \\ \mu_{(2)} & t > T_b \end{cases},$$

and $t = 1, 2, \ldots, T, T+1$ with $T = 50, 100, 200$. We set $b = \{0.95, 0.9\}$, $\lambda = (\mu_{(1)} - \mu_{(2)})/\sigma_\varepsilon = \{0.5, 1, 2\}$.[12] We assume that T_b, λ and q, the ratio of the pre- and post-beak error variances, are unknown and have to be estimated.

The third model adds a regressor, such that

$$y_t = \beta_t x_t + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0,1),$$

where

$$\beta_t = \begin{cases} \beta_{(1)} & t \leq T_b \\ \beta_{(2)} & t > T_b \end{cases},$$

we set b and λ as in the second experiment. Regressors are generated as $x_t \sim iid\mathrm{N}(0,1)$, and forecasts are conditional on x_{T+1}.

Forecasts based on the full estimation window with equal weights will serve as the base line to which all other forecasting methods are compared. We also include the infeasible optimal forecasts based on the optimal weights that use the true parameter values of the break process for comparison. For model (50) with continuous breaks the weights are given in (3), for the model with discrete breaks the weights are given in (12) and (13), and for the simple regression model they are given in (24) and (25).

[12] We also conducted experiments with a break in the error variance. As predicted by our theory, the results are qualitatively the same. For this reason they are omitted here, but can be found in the web supplement B.8.

For the first two models, we estimate γ from an MA(1) in first differences for the methods that assume a continuous break. We forecast the model with weights (3) using the estimated $\hat{\delta}$ and by ExpS with weights (9) using $\hat{\gamma}$.

For the linear regression model we replace ExpS with Constant Gain Least Squares (CGLS), which is a multivariate generalization of ExpS that has recently received increasing attention in the macroeconomic learning literature, see Evans and Honkapohja (2001) for a review and Markiewicz (2012) for a recent application. Under the CGLS approach the parameter vector, $\boldsymbol{\beta}_t$, is estimated by the following recursion

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \alpha \mathbf{R}_{t-1}^{-1} \mathbf{x}_t (y_t - \mathbf{x}_t' \boldsymbol{\beta}_{t-1}) \quad \text{and} \quad \mathbf{R}_t = \mathbf{R}_{t-1} + \alpha(\mathbf{x}_t \mathbf{x}_t' - \mathbf{R}_{t-1}), \tag{51}$$

where α is the downweighting parameter (also known as the forgetting factor) . Branch and Evans (2006) compare the forecasts of CGLS to other methods and find that it generally performs well. Similar to Branch and Evans (2006), we determine α using cross-validation. We calculate pseudo out-of-sample forecasts for the last 25 observations in the sample under consideration and choose the α that minimizes the MSFE over the values $0.01, 0.02, \ldots, 0.99$. Note that in the case where $\mathbf{x}_t = 1$ the above recursion reduces to the ExpS defined in (10) with $\alpha = 1 - \gamma$.[13]

For the methods that assume a discrete break process we use the Bai and Perron (1998, 2003) procedure to estimate the break dates, $\mathbf{b} = (b_1, b_2)'$, and, conditional on these estimates, the break sizes, $\boldsymbol{\lambda} = (\lambda_{(1)}, \lambda_{(2)})'$. We then use these estimates to compute feasible forecasts based on the optimal weights (12) and (13) in the random walk model or (24) and (25) in the simple linear regression model with $\hat{\mathbf{b}}$ and $\hat{\boldsymbol{\lambda}}$ in place of $\mathbf{b}$ and $\boldsymbol{\lambda}$. For the DGP with continuous breaks we allow for two breaks, for the DGP with a discrete break we restrict attention to testing for one break.

We also compute forecasts using the robust optimal weights developed in section 3. First, we assume that the forecaster uses the information that the break is in the last quarter but not in the last 2% of the sample. The corresponding weights are given by (44). Second, we assume that break dates in the full sample are equally likely with the weights given in (48). Finally, in the experiments with continuous break process we use robust optimal weights assuming two breaks, where the weights are calculated numerically integrating over b_1 and b_2 and $\phi_{(1)}$ and $\phi_{(2)}$ over the range -2 to 2.

For comparison, we construct forecasts based on the observations after the estimated break dates and forecasts based on the optimal estimation windows using the estimated break dates and sizes. Given the uncertainty over the break dates, we also average over different estimation windows, starting with the minimum window given by 5% of the available sample, namely we set $v_{\min} = 0.05$.

We construct one-period ahead forecasts for each method and base comparisons on the MSFE. We report ratios of MSFEs relative to that of the forecasts using equal weights, MSFE_{equal}, so that for method i we have

$$\text{rMSFE}_i = \frac{\text{MSFE}_i}{\text{MSFE}_{equal}}. \tag{52}$$

The results are based on 10,000 replications.

4.2 Monte Carlo Results

Continuous breaks DGP Table 2 reports the results for the DGP with continuous breaks. The first line contains the results for the infeasible optimal weights forecasts

[13] Further details of CGLS are provided in the web supplement B.10.

based on the true δ. Not surprisingly, these forecasts represent large improvements in MSFE relative to the equal weights forecasts.

The second and third lines contain the results for the optimal and ExpS weights forecasts using the estimates, $\hat{\delta}$ and $\hat{\gamma}$. As suggested in Section 2.1, these two forecasts are numerically identical for the parameters considered here. The forecast performance increases in T as γ is estimated more precisely for larger T.[14]

Amongst the methods that assume a discrete break, the robust optimal weights generally perform best. Notably, for a range of values of γ and sample sizes, they deliver the best forecasts of all feasible forecasts, including those based on the assumption of continuous breaks. In fact, only for $\gamma = 0.8$ and $T = 100$ and 200 do the optimal weights based on continuous breaks have the most precise forecasts.

Forecasts based on optimal weights under the assumption of discrete breaks perform well when $\gamma = 0.8$, that is when β_t has a large variation. For larger γ, the performance deteriorates and often does not improve on the equal weights forecasts. The results for the forecast from the optimal window are similar to the optimal weights forecast. The post-break window forecast is the least favorable forecasting method in this setting and often leads to the highest MSFEs.

Finally, the AveW forecasts performs well for a range of γ and sample sizes. Only when the true value of γ is small, does the AveW procedure perform less well since it does not discount past observations heavily enough.

Discrete breaks DGP Table 3 contains the results when the break in the drift of the underlying random walk process is discrete. The relative performance of the feasible forecasts primarily depends on the size of the break, and the sample size. The second line reports the results for the estimated optimal weights. When the break size, λ, is small, the detection of the break is difficult. As a consequence, the forecasts that use *estimated* optimal weights lead to higher MSFEs than most other forecasting methods. However, when $\lambda = 2$ the estimated optimal weights produce MSFEs that are among the smallest across all feasible methods. The benefit of applying optimal weights therefore depends on the ability to detect the break accurately.

The next two lines report the results for the robust optimal weights. For $\lambda = 0.5$ and 1, the forecasts that use the information that the break is in the last quarter of the sample provide the best forecasts across all feasible methods. The robust optimal weights that integrate b over the last quarter of the sample always perform better - and for larger breaks substantially so - than the robust optimal weights obtained by integrating over the entire sample, which shows how powerful this additional information is for the resulting forecasts. For large values of λ the robust optimal weights still improve vastly over the equal weights forecast but not as much as the estimated optimal weights.

Forecasts based on post-break observations (with estimated break dates) have the highest MSFE when the break size, λ, is small. Their performance improves dramatically when λ is large: the post-break forecasts have MSFEs very similar to the ones obtained for the estimated optimal weights. The optimal window forecasts perform quite similarly to the ones based on estimated optimal weights, and their performance depends largely on the size of the break.

AveW forecasts perform well when $T = 50$ and the break is small but less well for larger breaks. Still, in all examples, AveW offers substantial improvements over the

[14]MSE results for γ are available in web supplement B.8.

Table 2: Monte Carlo results for the random walk model with continuous breaks

	γ	0.8	0.9	0.95	0.98
	δ	0.224	0.105	0.051	0.020
$T = 50$					
opt.weight(cont.break; δ)		0.643	0.914	0.990	1.000
estim.opt.weight(cont.break; $\hat{\delta}$)		0.702	0.970	1.037	1.016
ExpS($\hat{\gamma}$)		0.702	0.970	1.037	1.016
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		0.723	1.031	1.106	1.112
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.645	0.967	1.093	1.119
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.738	0.921	0.996	1.019
rob.opt.weights(two breaks)		0.836	0.945	0.991	1.004
post-break obs.($\hat{b}$)		0.729	1.049	1.130	1.137
opt.window($\hat{b}, \hat{\lambda}$)		0.705	0.990	1.064	1.073
AveW($w_{\min} = 0.05$)		0.756	0.924	0.993	1.015
$T = 100$					
opt.weight(cont.break; δ)		0.458	0.777	0.949	0.996
estim.opt.weight(cont.break; $\hat{\delta}$)		0.469	0.804	0.974	1.014
ExpS($\hat{\gamma}$)		0.469	0.804	0.974	1.014
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		0.526	0.854	1.058	1.107
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.471	0.781	0.988	1.052
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.626	0.829	0.953	0.999
rob.opt.weights(two breaks)		0.764	0.889	0.966	0.996
post-break obs.($\hat{b}$)		0.528	0.862	1.076	1.128
opt.window($\hat{b}, \hat{\lambda}$)		0.520	0.828	1.018	1.066
AveW($w_{\min} = 0.05$)		0.649	0.840	0.955	0.998
$T = 200$					
opt.weight(cont.break; δ)		0.286	0.618	0.878	0.984
estim.opt.weight(cont.break; $\hat{\delta}$)		0.290	0.627	0.890	1.002
ExpS($\hat{\gamma}$)		0.290	0.627	0.890	1.002
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		0.363	0.682	0.959	1.077
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.326	0.625	0.881	1.010
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.530	0.739	0.907	0.985
rob.opt.weights(two breaks)		0.698	0.830	0.937	0.988
post-break obs.($\hat{b}$)		0.363	0.684	0.969	1.094
opt.window($\hat{b}, \hat{\lambda}$)		0.364	0.671	0.932	1.047
AveW($w_{\min} = 0.05$)		0.560	0.755	0.913	0.985

Note: The table reports the ratio of MSFE of forecasting method i relative to that using equal weights, $\text{MSFE}_i/\text{MSFE}_{equal}$. The DGP is $y_t = \beta_t + \sigma_\varepsilon \varepsilon_t$ where $\beta_t = \beta_{t-1} + \sigma_v v_t$, $\delta = \sigma_\varepsilon/\sigma_v$, and $\delta = (1-\gamma)/\sqrt{\gamma}$. Forecasting methods: (i) infeasible optimal weights as function of δ, (ii) optimal weights for continuous breaks where δ is estimated from an MA(1) in the first difference of the data, (iii) ExpS with γ estimated from an MA(1) in the first difference of the data, (iv) optimal weights based on point estimates of b and λ for up to two breaks, (v) robust optimal weights (44) with $\underline{b} = 0.75$ and $\bar{b} = 0.98$, (vi) robust optimal weights (48), (vii) robust optimal weights for two breaks with $\phi_{(1)}, \phi_{(2)} \in (-2, 2)$, (viii) post-break window based on $\hat{b}$, (ix) optimal window based on point estimates of b and λ for the last break, (x) AveW forecasts with $m = T(1-v_{\min})+1$ windows and $v_{\min} = 0.05$. The results are based on $R = 10,000$ repetitions.

Table 3: Monte Carlo results for random walk model with a discrete break in drift, $q = 1$

	b	0.95			0.9		
	λ	0.5	1	2	0.5	1	2
$T = 50$							
opt.weight(disc.break; b, λ)		0.923	0.653	0.284	0.910	0.634	0.276
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		1.040	0.873	0.428	1.040	0.842	0.342
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.955	0.708	0.443	0.940	0.656	0.334
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.956	0.857	0.751	0.940	0.810	0.662
post-break obs.($\hat{b}$)		1.060	0.885	0.427	1.060	0.856	0.343
opt.window($\hat{b}, \hat{\lambda}$)		1.004	0.847	0.451	1.003	0.813	0.349
AveW($w_{\min} = 0.05$)		0.966	0.888	0.805	0.948	0.836	0.709
estim.opt.weight(cont.break; $\hat{\delta}$)		0.994	0.961	0.798	0.992	0.930	0.577
ExpS($\hat{\gamma}$)		0.994	0.961	0.798	0.992	0.930	0.577
$T = 100$							
opt.weight(disc.break; b, λ)		0.893	0.603	0.256	0.875	0.592	0.257
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		1.022	0.826	0.320	1.014	0.737	0.263
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.912	0.701	0.473	0.884	0.619	0.316
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.953	0.867	0.775	0.931	0.805	0.662
post-break obs.($\hat{b}$)		1.039	0.839	0.319	1.030	0.747	0.262
opt.window($\hat{b}, \hat{\lambda}$)		0.991	0.800	0.329	0.986	0.722	0.268
AveW($w_{\min} = 0.05$)		0.965	0.900	0.830	0.940	0.831	0.706
estim.opt.weight(cont.break; $\hat{\delta}$)		0.992	0.944	0.666	0.984	0.847	0.337
ExpS($\hat{\gamma}$)		0.992	0.944	0.666	0.984	0.847	0.337
$T = 200$							
opt.weight(disc.break; b, λ)		0.869	0.571	0.238	0.862	0.577	0.248
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		1.010	0.711	0.245	0.984	0.618	0.249
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.894	0.685	0.461	0.867	0.605	0.306
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.949	0.863	0.771	0.928	0.802	0.658
post-break obs.($\hat{b}$)		1.027	0.720	0.244	0.998	0.621	0.249
opt.window($\hat{b}, \hat{\lambda}$)		0.984	0.695	0.249	0.966	0.613	0.251
AveW($w_{\min} = 0.05$)		0.962	0.899	0.831	0.937	0.828	0.704
estim.opt.weight(cont.break; $\hat{\delta}$)		0.989	0.898	0.391	0.973	0.727	0.265
ExpS($\hat{\gamma}$)		0.989	0.898	0.391	0.973	0.727	0.265

Note: The table reports the relative MSFEs for the DGP $y_t = \beta_t + \sigma_t \varepsilon_t$ with a break in β_t and σ_t at T_b. Here, $q = \sigma_{(1)}/\sigma_{(2)} = 1$. The first forecasting method uses optimal weights for a discrete break with known b and λ. For the remaining forecasting methods see Table 2.

full sample equal weights forecasts.

Forecasts that incorrectly assume the break process is continuous also reduce the MSFE relative to the full sample based forecasts but, as to be expected, are generally less efficient than those based on weights derived assuming a discrete break DGP. However, as T increases the forecasts of these methods improve considerably.

Table 4 reports the results for the simple linear regression model. While the magnitude of the relative MSFEs are affected by the additional variation introduced by the regressor, the relative ranking of the various forecasting methods is very similar to that for the random walk model. A notable difference is that the robust optimal weights now also deliver the best forecasts for the largest breaks when $T = 50$. CGLS

Table 4: Monte Carlo results for a single regressor and a discrete break, $q = 1$

	b	0.95			0.9	
	λ	0.5	1	2	0.5	1	2
$T = 50$							
opt.weight(disc.break; b, λ)		0.979	0.853	0.542	0.971	0.832	0.520
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		1.005	0.978	0.851	1.009	0.952	0.631
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.995	0.876	0.660	0.990	0.846	0.576
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.980	0.925	0.836	0.975	0.907	0.783
post-break obs.($\hat{b}$)		1.007	0.980	0.849	1.012	0.957	0.633
opt.window($\hat{b}, \hat{\lambda}$)		1.007	0.980	0.850	1.012	0.957	0.634
AveW($w_{\min} = 0.05$)		0.982	0.933	0.854	0.977	0.911	0.794
CGLS($\hat{\alpha}$)		1.306	1.144	0.985	1.295	1.087	0.818
$T = 100$							
opt.weight(disc.break; b, λ)		0.961	0.800	0.499	0.952	0.796	0.502
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		1.003	0.913	0.607	1.003	0.877	0.520
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.970	0.853	0.668	0.957	0.813	0.557
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.979	0.929	0.854	0.972	0.903	0.783
post-break obs.($\hat{b}$)		1.006	0.916	0.608	1.006	0.881	0.520
opt.window($\hat{b}, \hat{\lambda}$)		1.006	0.916	0.608	1.006	0.881	0.520
AveW($w_{\min} = 0.05$)		0.983	0.941	0.880	0.974	0.911	0.800
CGLS($\hat{\alpha}$)		1.118	0.999	0.774	1.094	0.911	0.590
$T = 200$							
opt.weight(disc.break; b, λ)		0.955	0.786	0.473	0.945	0.793	0.485
estim.opt.weight(disc.break; $\hat{b}, \hat{\lambda}$)		1.013	0.874	0.491	1.001	0.822	0.487
rob.opt.weights($\underline{b} = 0.75, \bar{b} = 0.98$)		0.963	0.846	0.637	0.949	0.809	0.533
rob.opt.weights($\underline{b} = 0, \bar{b} = 1$)		0.980	0.930	0.842	0.970	0.903	0.771
post-break obs.($\hat{b}$)		1.018	0.878	0.491	1.006	0.823	0.486
opt.window($\hat{b}, \hat{\lambda}$)		1.018	0.878	0.491	1.006	0.823	0.486
AveW($w_{\min} = 0.05$)		0.984	0.945	0.878	0.973	0.913	0.796
CGLS($\hat{\alpha}$)		1.017	0.892	0.559	1.001	0.854	0.560

Note: The results are for the simple linear regression model, $y_t = \beta_t x_t + \sigma_t \varepsilon_t$ with a single break in β_t at T_b. For definitions and forecasting procedures see the footnotes of Tables 2 and 3.

does relatively well for large breaks but is dominated by the optimal weights, which is not surprising, given that CGLS incorrectly assumes a continuous break process.

Overall, the Monte Carlo results suggest that when the break size is small and/or the sample is too small for an accurate estimation of the break process, the robust optimal weights developed in this paper deliver the most precise forecasts. This is true for discrete as well as continuous break processes. When the break process is continuous, the sample is sufficiently large, and γ not too close to unity, estimated optimal weights and ExpS forecasts with estimated down-weighting parameter will result in the most precise forecasts. If the true γ is large, robust optimal weights forecasts dominate even in large samples. Under discrete breaks that are large and easily identified, the optimal weight forecasts provide the best forecasts, otherwise robust optimal weights forecasts dominate.

5 Application to the yield curve as a predictor of real economic activity

5.1 The empirical model

The slope of the yield curve has emerged as a valuable leading indicator of GDP growth; see Stock and Watson (2003) for a survey of the literature. However, recent evidence suggests that the relationship between GDP growth and the yield curve may be subject to structural breaks (Estrella, Rodrigues and Schich 2003, Giacomini and Rossi 2006, Schrimpf and Wang 2010). We will use the forecasting methods discussed in the previous sections to investigate whether they can improve the forecasts of GDP growth with the slope of the yield curve as the predictor.

The forecasts are based on the regression model

$$y_{t,t+h} = \beta_0 + \beta_1 s_t + \varepsilon_t, \tag{53}$$

where $y_{t,t+h} = 100 \ln(Y_{t+h}/Y_t)$, Y_t is the level of real GDP at time t, and $s_t = i_t^L - i_t^S$ is the slope of the yield curve defined as the difference between the long term interest rate, i_t^L, and the short term interest rate, i_t^S. This specification is the most common in the literature (e.g., Estrella and Hardouvelis 1991, Estrella and Mishkin 1997, and the literature cited above).

We evaluate the forecasts for horizons $h = 1, 2, 3, 4$ quarters. An issue involving direct forecasts with horizons greater than one is the overlap implicit in the regressions. Pesaran, Pick and Timmermann (2011) show that accounting for the overlap of observations can lead to gains in forecast accuracy but that these gains materialize at forecast horizons that are larger than those considered here. In order not to complicate the forecast exercise further, we restrict attention to estimations that do not account for overlap.

The source of quarterly observations on GDP, long and short term interest rates is the data set (2009 vintage) available with the GVAR toolbox (Smith and Galesi 2012). The data set contains quarterly observations for 33 countries. As not all countries have long term bond markets, we focus on the following nine industrialized economies: Australia, Canada, France, Germany, Italy, Japan, Spain, UK, and USA. The data set start in 1979Q1 and ends in 2009Q4. Recursive out-of-sample forecasts are constructed, with the first forecast using the observations up to 1993Q4 for estimation.

We report results for the entire forecast period and for the sub-periods 1994Q1–2000Q4, 2001Q1–2006Q4, and 2007Q1-2009Q4. The first period includes the build-up of the dot-com bubble, the second contains the time after the dot-com bubble burst and the build-up of the sub-prime mortgage market, the third contains the observations following the collapse of the sub-prime mortgage market.

We will use the forecasting methods outlined in Section 4. However, we do not impose knowledge of the timing of the structural break on the optimal weights as such knowledge may not be available to the researcher at the time. Given that we have more than one regressor, we use the optimal weights (38) and (39), where we estimate $\mathbf{\Omega}_{xx}$ over the estimation sample available for the forecast. For the CGLS forecasts, we estimate α in (51) via cross-validation for each forecast horizon separately. We use the last 40 observations of the presample up to 1993Q4 to obtain quasi-out-of-sample forecasts and choose the α that minimizes the MSFE over the values $0.01, 0.02, \ldots, 0.99$ (the estimates for α are available in web supplement B.10). Cross-validation relies on the assumption that the underlying model does not change in the forecast period.

Given the uncertainty that surrounds the estimates of the downweighting parameter, we also generate forecasts using a cross-country average of the downweighting estimates $\bar{\hat{\alpha}}(h) = \frac{1}{m}\sum_{i=1}^{m}\hat{\alpha}_i(h)$, where $\hat{\alpha}_i(h)$ is the estimate of the downweighting parameter for the i^{th} country when the forecast horizon is h.[15]

Forecasts are compared using the relative MSFE measure defined in (52). Furthermore, we test for equal forecast performance using a generalization of the panel version of the Diebold-Mariano test proposed by Pesaran, Schuermann and Smith (2009), where we allow for $h > 1$ and general cross-country aggregation weights. Details of the test are provided in Appendix A.2.

5.2 Results for GDP growth forecasts

Table 5 give cross-country averages of MSFEs over the full sample, and over the three sub-samples. In the aggregation of individual country MSFEs we use both GDP-PPP based weights as well as equal weights. The table also shows if a forecasting method outperforms the equal weights forecast significantly using the panel version of the Diebold and Mariano test statistic. An asterisk denotes forecasts that are significantly better than the equal weights forecast at the 5% significance level.

The first line of the table shows average MSFEs for the equal weights forecasts. The second line gives the relative MSFE of the forecasts using optimal weights based on the estimated break date and size. For both country aggregation schemes, the estimated optimal weights forecasts improve on equal weights forecasts for $h = 1$ but not for larger horizons. In general, while the optimal weights forecasts have a lower MSFE than the post-break and CGLS forecasts, they are less precise than the remaining forecasting methods.

Forecasts using robust optimal weights, in contrast, deliver vastly improved forecasts compared to equal weights. They provide the best forecasts for all horizons, except for $h = 1$ and when relative GDP weights are used to aggregate the MSFEs across countries. Only in this case the CGLS forecast has a lower MSFE than the robust optimal weight forecast. Also, the improvements over equal weights are statistically significant for all horizons. While the MSFEs of the robust optimal weights for one break are generally smaller than that for robust optimal weights for two breaks, the latter's improvements are significant over all horizons, whereas over the entire forecast period the formers are only significant for $h = 1$ and 2.

Post-break window forecasts are substantially worse than equal weights forecasts. AveW forecasts, in contrast, improve over the equal weights forecasts and the improvement is statistically significant for $h = 1$, no matter which cross country aggregation scheme is used, and for $h = 2$ if the countries are weighted equally in the averaging process. Finally, CGLS with country specific estimates of α delivers the least precise forecasts. Pooling the estimates of α across countries leads to results that are significantly better than the equal weighted forecasts for $h = 1$. For larger h, however, pooled α CGLS forecasts fail to improve on equal weights forecasts.

When considering the sub-samples separately some interesting additional patterns emerge. We note that in the first sub-sample (covering forecasts for the period 1994Q1–2000Q4) robust optimal weights for one and two breaks deliver the best forecasts for all forecast horizons irrespective of how the country results are aggregated, and these forecast improvements are statistically significant at all horizons. The second sub-

[15] In practice, prior knowledge of α may be available but, following the suggestion of an anonymous referee, we make no such assumption in this forecasting exercise.

Table 5: Predictive power of the yield curve: Relative forecast accuracy averaged across countries

	GDP weighted ave.				Equally weighted ave.			
h	1	2	3	4	1	2	3	4
All forecasts: 1994Q1–2009Q4								
equal weight(MSFE)	0.557	1.726	3.339	5.215	0.521	1.585	3.046	4.736
est.asy.opt.weight	0.947	1.026	1.052	1.048	0.974	1.082	1.087	1.069
rob.weight(1 break)	0.895*	0.900*	0.932	0.970	0.915*	0.946*	0.973	0.993
rob.weight(2 breaks)	0.942*	0.942*	0.954*	0.973*	0.953*	0.963*	0.974*	0.980*
post-break	1.111	1.131	1.118	1.076	1.191	1.154	1.133	1.074
AveW	0.994*	0.999	0.998	1.008	0.991*	0.996*	0.997	1.000
CGLS($\hat{\alpha}$)	2.612	1.192	1.305	1.298	4.389	1.456	1.336	1.507
CGLS($\bar{\hat{\alpha}}$)	0.878*	1.154	1.161	1.454	0.924*	1.392	1.365	2.024
Subsample 1: 1994Q1–2000Q4								
equal weight(MSFE)	0.374	1.034	2.066	3.612	0.352	0.908	1.753	2.970
est.asy.opt.weight	1.009	0.972	1.139	1.062	1.018	1.080	1.344	1.070
rob.weight(1 break)	0.855*	0.786*	0.774*	0.797*	0.911*	0.925*	0.943*	0.965*
rob.weight(2 breaks)	0.928*	0.894*	0.891*	0.905*	0.956*	0.956*	0.960*	0.970*
post-break	1.025	1.047	1.045	1.192	1.060	1.185	1.228	1.183
AveW	0.994*	1.013	1.011	1.004*	0.993*	1.009	1.023	1.028
CGLS($\hat{\alpha}$)	0.976*	0.909*	0.954	1.488	1.266	1.332	1.446	2.671
CGLS($\bar{\hat{\alpha}}$)	0.832*	0.776*	1.141	2.538	0.893*	1.039	1.953	5.282
Subsample 2: 2001Q1–2006Q4								
equal weight(MSFE)	0.230	0.633	1.115	1.576	0.198	0.547	0.971	1.374
est.asy.opt.weight	1.009	1.072	1.046	1.229	1.000	1.145	1.080	1.356
rob.weight(1 break)	0.989	1.066	1.231	1.322	0.962*	0.996	1.058	1.101
rob.weight(2 breaks)	0.984*	0.996	1.042	1.053	0.974*	0.975	0.990	0.994
post-break	1.026	1.113	1.074	1.189	1.021	1.241	1.129	1.329
AveW	0.988*	0.983*	0.982*	0.986	0.980*	0.973*	0.969*	0.972*
CGLS($\hat{\alpha}$)	1.283	1.631	3.471	3.401	2.129	1.624	2.403	2.874
CGLS($\bar{\hat{\alpha}}$)	1.064	1.410	2.247	3.084	1.051	1.451	2.002	2.750
Subsample 3: 2007Q1–2009Q4								
equal weight(MSFE)	1.636	5.471	10.542	15.832	1.559	5.185	9.998	15.137
est.asy.opt.weight	0.895	1.042	1.061	1.001	0.949	1.068	1.030	1.003
rob.weight(1 break)	0.889*	0.915*	0.948	0.978	0.911*	0.954*	0.983	1.007
rob.weight(2 breaks)	0.938*	0.953*	0.968*	0.978*	0.949*	0.969*	0.983*	0.994*
post-break	1.166	1.196	1.196	1.000	1.281	1.148	1.127	0.999
AveW	0.996*	0.997	0.998	1.002	0.995*	0.998	0.998	1.000
CGLS($\hat{\alpha}$)	3.163	1.166	1.034	0.961	5.649	1.400	1.206	1.058
CGLS($\bar{\hat{\alpha}}$)	0.848	1.213	1.017	0.912	0.932	1.488	1.159	1.032

Note: The table reports MSFEs of the forecasts with equal weights and for all other forecasting methods the ratio of MSFEs, that is, the MSFE of forecasting method i relative to that using equal weights, $\text{MSFE}_i/\text{MSFE}_{\text{equal}}$, for different forecast horizons, h. Forecasting methods: (i) equal weights, (ii) optimal weights (38) and (39) for discrete breaks based on point estimates of b and λ for up to five breaks, and $\mathbf{\Omega}_{xx}$ is estimated over the sample available for each forecast, (iii) robust optimal weights (48) that integrate the break date over the entire sample, (iv) robust optimal weights for two breaks with $\phi_{(1)}, \phi_{(2)} \in (-2, 2)$, (v) post-break window, (vi) AveW forecasts of Section 2.2.2 with $m = T(1 - v_{\min}) + 1$ windows and $v_{\min} = 0.05$, (vii) constant gain least squares estimates with the gain parameter α estimated using a cross-validation procedure (see web supplement B.10 for details). The dates given above denote the periods for which one-period ahead forecasts are made. The $h = 2$ forecast makes the first forecast for the observation one quarter later, the $h = 3$ forecast for that two periods later, and the $h = 4$ forecast for that three quarters later. The GDP weighted average uses weights $w_i = Y_i/(\sum_{j=1}^{m} Y_j)$, where Y_i is the 2008 GDP in purchasing power terms for country i available from the GVAR data base and $m = 9$ is the number of countries. The equal weights average uses $w_i = 1/m$. An asterisk denotes forecast that is significantly better than that obtained from equal weights according to the panel Diebold-Mariano test statistic at a 5% significance level.

sample (2001Q1–2006Q4) offers a different picture. Most forecasting methods cannot improve on the equal weights forecasts. The exceptions are the AveW forecasts and, for $h = 1$, the robust optimal weights forecasts. AveW is the only forecasting method that delivers significant improvements irrespective of horizon and country weights. Robust optimal weights improve on equal weights, too, but the difference is not statistically significant for $h > 1$. In the third sub-sample (2007Q1–2009Q4) GDP growth is much harder to forecast as indicated by the MSFE of the equal weights forecasts. Forecasts based on robust optimal weights can improve the forecast by over 10%, and those for two breaks deliver significant improvements for all horizons. The relative performance is similar to that of the first sub-period: the robust optimal weights provide the best results, whereas forecasts that require estimates of break dates perform poorly. AveW forecasts deliver modest improvements. CGLS, in contrast, generally leads to worse forecasts.

Overall, forecasting methods that rely on estimates of break dates perform poorly in this application. AveW leads to modest but consistent improvements over equal weights forecasts. Robust optimal weights forecasts lead to large and, in the majority of cases, statistically significant improvements over equal weights forecasts.

6 Conclusion

This paper presents a new approach to forecasting in the presence of structural breaks. Under continuous break processes our approach approximates the exponential smoothing weights that have long been considered in the literature. Under discrete breaks, our approach delivers new forecasts based on optimal weights. In practice, dates and sizes of breaks are unknown and their estimates can be unreliable. For such cases we derive robust optimal weights, (46) and (47), that do not require *a priori* knowledge of the break dates or their sizes. Should information about the range of a break point be available, for example, from the confidence interval of a break point test, this can be incorporated in the robust optimal weights (44).

We evaluate the forecasting performance of the different weighting schemes in Monte Carlo experiments and in an application to forecasts of GDP growth across nine industrialized economies using the slope of the yield curve as a predictor. Forecasts based on robust optimal weights, which require neither knowledge of the break dates nor a downweighting parameter, lead to forecasts that perform better than other feasible alternatives in a wide range of settings.

A Appendix: Mathematical details

A.1 Derivation of optimal weights for multiple regression model with a single break

Using $\hat{\boldsymbol{\beta}}_T(\mathbf{w})$ in (19) we have

$$\hat{\boldsymbol{\beta}}_T(\mathbf{w}) - \boldsymbol{\beta}_{(2)} = \mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})(\boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(2)}) + \mathbf{S}^{-1}(\mathbf{w})\sum_{t=1}^{T} w_t\mathbf{x}_t\sigma_t\varepsilon_t.$$

Hence,

$$\begin{aligned} e_{T+1}(\mathbf{w}) &= \mathbf{y}_{T+1} - \mathbf{x}'_{T+1}\hat{\boldsymbol{\beta}}_T(\mathbf{w}) = -\mathbf{x}'_{T+1}\left[\hat{\boldsymbol{\beta}}_T(\mathbf{w}) - \boldsymbol{\beta}_{(2)}\right] + \sigma\varepsilon_{T+1}, \\ &= \sigma_{T+1}\varepsilon_{T+1} - \mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})(\boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(2)}) - \mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\sum_{t=1}^{T} w_t\mathbf{x}_t\sigma_t\varepsilon_t. \end{aligned}$$

Dividing by $\sigma^2_{(2)}$ and taking expectations of the squared forecast error yields (20).

In order to obtain the optimal weights we minimize (20) with respect to $\mathbf{w}$ subject to $\boldsymbol{\iota}'_T\mathbf{w} = 1$. Defining θ as the Lagrange multiplier associated with $\boldsymbol{\iota}'_T\mathbf{w} = 1$, the first order conditions for the above optimization problem are given by the following. For $t \leq T_b$

$$\begin{aligned} &\left[q^2\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{T+1}\right] w_t \\ &\quad = \theta/2 + \left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]\left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right] \\ &\qquad + \mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w})\left(\sum_{t=1}^{T_b} q^2 w_t^2\mathbf{x}_t\mathbf{x}'_t + \sum_{t=T_b+1}^{T} w_t^2\mathbf{x}_t\mathbf{x}'_t\right)\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{T+1} \\ &\qquad - \left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]\left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\boldsymbol{\lambda}\right], \end{aligned}$$

where $\mathbf{A}_t = \mathbf{x}_t\mathbf{x}'_t$ and for $t \geq T_b + 1$

$$\begin{aligned} &\left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{T+1}\right] w_t \\ &\quad = \theta/2 + \left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]\left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right] \\ &\qquad + \mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{A}_t\mathbf{S}^{-1}(\mathbf{w})\left(\sum_{t=1}^{T_b} q^2 w_t^2\mathbf{x}_t\mathbf{x}'_t + \sum_{t=T_b+1}^{T} w_t^2\mathbf{x}_t\mathbf{x}'_t\right)\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{T+1}. \end{aligned}$$

Multiplying both sides of the above two expressions by w_t and aggregating across $t = 1, 2, \ldots, T$ it is again easily seen that $\theta = 0$.

If $\mathbf{A}_t = 0$ the solution for w_t is indeterminate and without loss of generality can be set to 0. So we consider solutions where $\mathbf{A}_t \neq 0$, which yields for $t \leq T_b$

$$\begin{aligned} w_t &= \frac{\left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]\left[\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]}{q^2\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\boldsymbol{x}_t} - \frac{\left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]\left[\mathbf{x}'_t\boldsymbol{\lambda}\right]}{q^2\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\boldsymbol{x}_t} \\ &\quad + \frac{\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\left(\sum_{t=1}^{T_b} q^2 w_t^2\mathbf{x}_t\mathbf{x}'_t + \sum_{t=T_b+1}^{T} w_t^2\mathbf{x}_t\mathbf{x}'_t\right)\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{T+1}}{q^2\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\boldsymbol{x}_t} \end{aligned} \tag{54}$$

and for $t \geq T_b + 1$

$$\begin{aligned} w_t &= \frac{\left[\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]\left[\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right]}{\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\boldsymbol{x}_t} \\ &\quad + \frac{\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\left(\sum_{t=1}^{T_b} q^2 w_t^2\mathbf{x}_t\mathbf{x}'_t + \sum_{t=T_b+1}^{T} w_t^2\mathbf{x}_t\mathbf{x}'_t\right)\mathbf{S}^{-1}(\mathbf{w})\mathbf{x}_{T+1}}{\mathbf{x}'_{T+1}\mathbf{S}^{-1}(\mathbf{w})\boldsymbol{x}_t}. \end{aligned} \tag{55}$$

The last result follows since $\left[\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\boldsymbol{\lambda}\right] - \left[\mathbf{x}'_t\boldsymbol{\lambda}\right]$ can be written as

$$-\mathbf{x}'_t\left[\mathbf{I}_k - \mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_1(\mathbf{w}_{(1)})\right]\boldsymbol{\lambda} = -\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\left[\mathbf{S}(\mathbf{w}) - \mathbf{S}_1(\mathbf{w}_{(1)})\right]\boldsymbol{\lambda} = -\left[\mathbf{x}'_t\mathbf{S}^{-1}(\mathbf{w})\mathbf{S}_2(\mathbf{w}_{(2)})\boldsymbol{\lambda}\right].$$

Rearranging (54) and (55) yields the results in (21) and (22).

A.2 Panel test of forecast performance

We evaluate the performance of different forecasting methods against the equal weights forecast using a generalization of the test proposed by Pesaran, Schuermann and Smith (2009). Our test allows for $h > 1$ and general aggregation weights across countries. Consider the quadratic loss differential $z_{it}(h) = [e_{it,A}(h)]^2 - [e_{it,B}(h)]^2$, where $e_{it,A}$ is the forecast error of method A and $e_{it,B}$ that of the benchmark forecast, $i = 1, 2, \ldots, m$ denotes different countries, $t = 1, 2, \ldots, n$ different forecasts, and h is the forecast horizon.

For each horizon h, the pooled DM statistic tests the null hypothesis $H_0 : \alpha_i(h) = 0$ against the alternative $H_1 : \alpha_i(h) < 0$, for some i, where $\alpha_i(h)$ is defined by $z_{it}(h) = \alpha_i(h) + u_{it}(h)$, $u_{it}(h) \sim \left[0, \sigma_i^2(h)\right]$, and $u_{it}(h)$ are independently distributed across i, but can be serially correlated over t if $h > 1$. For a given h and a set of country weights, $\omega = (\omega_1, \omega_2, \ldots, \omega_m)'$, the pooled DM test statistic is

$$PDM(h) = \frac{\bar{z}(h)}{\sqrt{V(\bar{z}(h))}},$$

where $\bar{z}(h) = \mathbf{w}'\bar{\mathbf{z}}(h)$, $\bar{\mathbf{z}}(h) = (\bar{z}_1(h), \bar{z}_2(h), \ldots, \bar{z}_m(h))'$, $\bar{z}_i(h) = \frac{1}{n}\sum_{t=1}^{n} z_{it}(h)$, and $V(\bar{z}) = \frac{1}{n}\omega'\mathbf{S}(h)\omega$, where $\mathbf{S}(h)$ is a diagonal matrix with $\sigma_i^2(h)$, the variance of $\bar{z}_i(h)$, as the (i,i) element on its diagonal. For $h = 1$, the variance, $\sigma_i^2(1)$, is estimated as

$$\hat{\sigma}_i^2(1) = \frac{1}{n-1}\sum_{t=1}^{n}\left[z_{it}(1) - \bar{z}_i(1)\right]^2.$$

For $h > 1$, we estimate $\sigma_i^2(h)$ nonparametrically to allow for the autocorrelation of $z_{it}(h)$ due to the overlap of the underlying forecast errors. Using a Bartlett window, we have

$$\hat{\sigma}_i^2(h) = \frac{1}{n-1}\sum_{t=1}^{n}\left[z_{it}(h) - \bar{z}_i(h)\right]^2 + \frac{2}{n}\sum_{j=1}^{s}(1 - \frac{j}{s+1})\sum_{t=j+1}^{n}\left[z_{it}(h) - \bar{z}_i(h)\right]\left[z_{i,t-j}(h) - \bar{z}_i(h)\right].$$

Under the null hypothesis where the serial correlation in $z_{it}(h)$ for $h > 1$ is solely due to overlap in the forecast errors $s = h - 1$, and this is our choice in the application. Under the null hypothesis of no relative forecasting skill, the *PDM* statistic is asymptotically distributed as $N(0, 1)$. Note that the test is set up as a one-sided test. Thus, the 5% critical value is 1.64.

References

Altissimo, Filippo, and Valentina Corradi (2003) 'Strong rules for detecting the number of breaks in a time series.' *Journal of Econometrics* 117, 207–244.

Andrews, Donald W. K. (1993) 'Tests for parameter instability and structural change with unknown change point.' *Econometrica* 61, 821–856.

Andrews, Donald W. K., Inpyo Lee and Werner Ploberger (1996) 'Optimal changepoint tests for normal linear regression.' *Journal of Econometrics* 70, 9–38.

Bai, Jushan (1997) 'Estimation of a change point in multiple regression models.' *Review of Economics and Statistics* 79, 551–563.

Bai, Jushan, and Pierre Perron (1998) 'Estimating and testing linear models with multiple structural changes.' *Econometrica* 66, 47–78.

Bai, Jushan, and Pierre Perron (2003) 'Computation and analysis of multiple structural change models.' *Journal of Applied Econometrics* 18, 1–22.

Branch, William, and George W. Evans (2006) 'A simple recursive forecasting model.' *Economics Letters* 91, 158–166.

Brown, Robert G. (1959) *Statistical Forecasting for Inventory Control.* New York: McGraw-Hill.

Brown, R. L., J. Durbin, and J. M. Evans (1975) 'Techniques for testing the constancy of regression relationships over time.' *Journal of the Royal Statistical Society B* 37, 149–192.

Clements, Michael P. and David F. Hendry (1999) *Forecasting Non-stationary Economic Time Series.* Cambridge, Mass.: MIT Press.

Clements, Michael P. and David F. Hendry (2006) 'Forecasting with breaks.' in G. Elliott, C.W.J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*. Elsevier, 605–657.

Estrella, Arturo, and Gikas A. Hardouvelis (1991) 'The term structure as a predictor of real economic activity.' *Journal of Finance* 46, 555–576.

Estrella, Arturo, and Frederic S. Mishkin (1997) 'The predictive power of the term structure of interest rates in Europe and the United States: Implications for the European Central Bank.' *European Economic Review* 41, 1375–1401.

Estrella, Arturo, Anthony P. Rodriguez, and Sebastian Schich (2003) 'How stable is the predictive power of the yield curve? Evidence from Germany and the United States.' *Review of Economics and Statistics* 85, 629–644.

Evans, George W., and Seppo Honkapohja (2001) *Learning and Expectations in Macroeconomics*. Princeton: Princeton University Press.

Giacomini, Raffaella, and Barbara Rossi (2006) 'How stable is the forecasting performance of the yield curve for output growth?' *Oxford Bulletin of Economic and Statistics* 68, 783–795.

Giacomini, Raffaella, and Barbara Rossi (2009) 'Detecting and predicting forecast breakdowns.' *Review of Economic Studies* 76, 669–705.

Hinkley, David V. (1970) 'Inference about the change-point in a sequence of random variables.' *Biometrika* 57, 1–17.

Holt, Charles (1957) 'Forecasting trends and seasonals by exponential weighted averages.' *ONR Memorandum* 52/1957, Carnegy Mellon University.

Hyndman, Rob J., Anna Koehler, J. Keith Ord, and Ralph D. Snyder (2008) *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer Verlag.

Inoue, Atsushi, and Barbara Rossi (2011) 'Identifying the sources of instabilities in macroeconomic fluctuations.' *Review of Economics and Statistics* 164, 158–172.

Koop, Gary, and Simon M. Potter (2007) 'Estimation and forecasting in models with multiple breaks.' *Review of Economic Studies* 74, 763–789.

Markiewicz, Agnieszka (2012) 'Model uncertainty and exchange rate volatility.' *International Economic Review* 53, 815–843.

Pesaran, M. Hashem, Davide Pettenuzzo, and Allan Timmermann (2006) 'Forecasting time series subject to multiple structural breaks.' *Review of Economic Studies* 73, 1057–1084.

Pesaran, M. Hashem, and Andreas Pick (2011) 'Forecast combination across estimation windows.' *Journal of Business and Economic Statistics* 29, 307–318.

Pesaran, M. Hashem, Andreas Pick, and Allan Timmermann (2011) 'Variable selection, estimation and inference for multi-period forecasting problems.' *Journal of Econometrics* 164, 173–187.

Pesaran, M. Hashem, Til Schuermann, and L. Vanessa Smith (2009) 'Forecasting economic and financial variables with global VARs.' *International Journal of Forecasting* 25, 642–675.

Pesaran, M. Hashem, and Allan Timmermann (2002) 'Market timing and return predictability under model instability.' *Journal of Empirical Finance* 9, 495–510.

Pesaran, M. Hashem, and Allan Timmermann (2005) 'Small sample properties of forecasts from autoregressive models under structural breaks.' *Journal of Econometrics* 129, 129–217.

Pesaran, M. Hashem, and Allan Timmermann (2007) 'Selection of estimation window in the presence of breaks.' *Journal of Econometrics* 137, 134–161.

Rossi, Barbara (2011) 'Advances in forecasting under instability.' Chapter prepared for the *Handbook of Economic Forecasting*, eds. G. Elliot and A. Timmermann.

Schrimpf, Andreas, and Qingwei Wang (2010) 'A reappraisal of the leading indicator properties of the yield curve under structural instability.' *International Journal of Forecasting* 26, 836–857.

Smith, L. V. and A. Galesi (2012), Global VAR Modelling, `https://sites.google.com/site/gvarmodelling/`.

Stock, James H., and Mark W. Watson (2003) 'Forecasting output and inflation: The role of asset prices.' *Journal of Economic Literature* 41, 788–829.