Empirical Economics (2022) 63:1–41
https://doi.org/10.1007/s00181-021-02137-w

Using structural break inference for forecasting time series

Gantungalag Altansukh[1] · **Denise R. Osborn**[2]

Received: 8 January 2021 / Accepted: 10 September 2021 / Published online: 10 November 2021

Abstract
Rather than relying on a potentially poor point estimate of a coefficient break date when forecasting, this paper proposes averaging forecasts over sub-samples indicated by a confidence interval or set for the break date. Further, we examine whether explicit consideration of a possible variance break and the use of a two-step methodology improves forecast accuracy compared with using heteroskedasticity robust inference. Our Monte Carlo results and empirical application to US productivity growth show that averaging using the likelihood ratio-based confidence set typically performs well in comparison with other methods, while two-step inference is particularly useful when a variance break occurs concurrently with or after any coefficient break.

Keywords Forecasting time series · Structural breaks · Confidence intervals · Combining forecasts · Productivity growth

JEL Classifications C32 · C53

1 Introduction

The pervasiveness of structural breaks in many macroeconomic time series is widely acknowledged (Stock and Watson 1996; Paye and Timmermann 2006) and they are an important source of a forecast failure (Hendry 2000; Hendry and Clements 2003). This paper considers a scenario in which a discrete and permanent change in model coefficients may occur during the sample period used for estimation. Various forecast methods and strategies are proposed in the literature to deal with such a possibility, and these can be broadly classified into those that employ an estimated break date and robust methods that treat the break date as unknown.

✉ Denise R. Osborn
denise.osborn@manchester.ac.uk

[1] Department of Economics, National University of Mongolia, Ulaanbaatar 11000, Mongolia

[2] Economics, School of Social Sciences, University of Manchester, Manchester M13 9PL, UK

When the timing of a break is known or estimated precisely, unbiased estimates of the coefficients can be obtained using observations after the break, leading to the post-break window forecast. However, even with the accurate estimate of a break, if the length of the post-break sample is relatively short, the post-break coefficients and size of the break may be poorly estimated. Although the use of pre-break observations introduces estimation bias, it is often optimal to include some pre-break observations in the estimation sample to reduce the forecast error variance, as shown analytically and empirically by Pesaran and Timmermann (2004, 2005, 2007). The forecast accuracy of such methods heavily relies on how well the true break date is estimated and, in practice, theoretical gains may not be fully exploited as estimates of break dates can be imprecise (Elliott 2005; Paye and Timmermann 2006). A different approach is taken by Inoue et al. (2017), who propose using a rolling window with the optimal estimation window size selected by minimizing the conditional mean square forecast error.

Rather than producing a forecast based on a single estimation window, recent findings suggest that forecast combination methods that average over a model estimated with different sizes of windows often produce more accurate forecasts (Pesaran and Pick 2011; Eklund et al. 2013; Tian and Anderson 2014; Koo and Seo 2015; Boot and Pick 2020). Such methods typically assume the date of a break is unknown, and hence distortions from imprecise break date estimates are generally mitigated by averaging. However, other factors play a role, with forecast combination methods working well when breaks are small, occur frequently, towards the end of the sample or affect only the variance, and can perform relatively poorly in the presence of large breaks (Pesaran and Timmermann 2007; Pesaran and Pick 2011; Eklund et al. 2013). On the other hand, large breaks are easier to detect and hence a carefully selected single estimation window based on break date information can be preferable for forecasting (Pesaran and Timmermann 2007; Pesaran et al. 2013).

These results suggest that information on the nature of a possible break is important in deciding whether to use such information when forecasting or to adopt forecast methods that are robust to a presence of a possible break. This paper proposes a simple but intuitive approach based on forecast combinations and explicitly using break date estimates. To be specific, we propose employing a confidence interval or confidence set for the estimated break date instead of relying on a point estimate which may be poorly identified. We treat each date in the confidence interval/set as one of a sequence of choices for the potential break date and the corresponding post-break window forecasts are averaged. This incorporates information on the size of a break, since a break that is large (relative to the sample size) implies a narrow interval. Our approach is designed to improve on existing robust methods that combine forecasts from all possible windows (Pesaran and Pick 2011), by excluding windows that use less relevant data and which can yield large forecast errors. Koo and Seo (2015) employ a similar approach in the context of a misspecified model, but our interest lies in the situation where the model is correctly specified.[1] We employ the break date confidence interval originally proposed by Bai (1997), which is widely employed in

[1] Koo and Seo (2015) argue against using a confidence set for the break date due to its poor coverage in their misspecified model. Our analysis also examines empirical coverage, which can be good when the break model is correctly specified (see Table 2).

the structural break methodology of Bai and Perron (1998), and the confidence set of Eo and Morley (2015). Furthermore, to shed light on the nature of a detected break and to gain efficiency in estimation, a stepwise testing approach for changes in individual coefficients can be incorporated into the method.

A second contribution of this paper lies in the consideration given to the role of variance breaks. Although these are often overlooked in the forecast literature, it is known that tests for the presence of coefficient breaks are affected when breaks can also occur in the disturbance variance (Bai and Perron 2006). To be explicit, we consider the situation where both the coefficients and disturbance variance may be subject to a single structural break during the sample period, but these two breaks do not necessarily occur concurrently. Under such circumstances, the forecaster may apply a heteroskedasticity consistent (HC) procedure to test for a coefficient break. Another option is to explicitly examine the possibility of variance change and we investigate whether taking this route by use of a two-step break point testing methodology can improve forecast accuracy compared with HC testing. The procedure we employ, which allows the possibility of distinct coefficient and variance breaks occurring during the estimation sample, is built on Bataa et al. (2013) and also adapted by Altansukh et al. (2017).

Forecast performance is assessed through both Monte Carlo simulations and an empirical application to US productivity growth series. Our confidence interval/set approach is compared with widely advocated forecast approaches, including post-break, trade-off, cross-validation and window averaging methods proposed by Pesaran and Timmermann (2007) and Pesaran and Pick (2011). The simulation results show that our method performs well when the confidence set of Eo and Morley (2015) is employed, regardless of the size and nature of a break, with the empirical application supporting its usefulness for improving forecast accuracy in the presence of structural change. Further, the two-step testing approach is also generally beneficial, and this is especially the case when either no coefficient break applies or when a coefficient break occurs concurrently with or prior to a variance break.

This paper proceeds as follows. Section 2 outlines the confidence interval forecast method and describes the structural break inference methods that we employ. Section 3 sets up the Monte Carlo simulations and the simulation results are presented in Sect. 4. Section 5 examines the performance of forecast methods for US productivity growth and Sect. 6 concludes.

2 Methodology

This section outlines the forecasting methods that we employ, focusing particularly on variance breaks and exploiting information contained in coefficient break date confidence intervals/sets.

2.1 Forecast methods

For forecasting purposes, consider the dynamic model

$$y_t = \beta_t' \mathbf{x}_{t-1} + \sigma_t \varepsilon_t \qquad \varepsilon_t \sim \text{IID}(0, 1) \tag{1}$$

where $\mathbf{x}_{t-1}$ is a $k \times 1$ vector of regressors whose values are known at time $t-1$, β_t is the $k \times 1$ coefficient vector for x_{t-1},while ε_t is an error term that is serially uncorrelated and uncorrelated with $\mathbf{x}_{t-1}$. The regressor vector $\mathbf{x}_{t-1}$ will typically include at least one lag of y_t, but in the absence of structural breaks in (1) $\mathbf{x}_{t-1}$ is covariance stationary. Using sample period data for $t = 1, \ldots, T$, our interest lies in forecasting future values of y when β_t and the disturbance variance σ_t^2 may each be subject to a single within sample structural break, with the two types of change not necessarily coinciding. In particular, y_{T+h} $(h = 1, 2, ..)$ is to be forecast using the observations[2] $\Gamma_T = \{\mathbf{x}_t : t = 1, 2, \ldots, T\}$, recognizing that, in practice, neither the occurrence of breaks nor the date(s) at which they occur are known. Our aim is to exploit structural break inference information to improve forecast accuracy.[3]

Ignoring any possible structural break(s), the full sample one-step ahead forecast is

$$\hat{y}_{T+1,\text{Base}} = \hat{\beta}_{1:T}' \, \mathbf{x}_T \tag{2}$$

where $\hat{\beta}_{1:T}$ is obtained by applying OLS estimation to (1) using all T sample observations. Even when breaks may occur, this full sample forecast provides a benchmark for assessing the performance of methods which allow the possibility of breaks. Now suppose that, by some appropriate method, a structural break in the coefficient vector is estimated to have occurred at $t = \hat{T}_c$, where $1 < \hat{T}_c < T$. With sufficient observations available to estimate the coefficient vector in the period after the estimated break, the usual post-break forecast is

$$\hat{y}_{T+1,PB} = \hat{\beta}_{\hat{T}_c+1:T}' \, \mathbf{x}_T \tag{3}$$

where $\hat{\beta}_{\hat{T}_c+1:T}$ is obtained by OLS using observations $t = \hat{T}_c + 1, \hat{T}_c + 2, \ldots, T$.

If variance breaks may be present, heteroskedastic consistent (HC) inference can be employed for coefficient break testing, with OLS then applied to (2) or (3) as appropriate. Although asymptotically valid, the simulation evidence of Bai and Perron (2006) and Pitarakis (2004) indicates that HC inference leads to over-sized coefficient break tests in finite samples. This can be serious for forecasting, because a false conclusion that a break exists leads to a reduced effective sample size for coefficient estimation, implying a loss of efficiency and an increase in theoretical mean square forecast error.

[2] We assume pre-sample observations are available such that (1) can be applied for $t = 1$.

[3] Tian and Anderson (2014) use information from the reverse ordered cusum test to provide forecast combination weights, but we base our analysis on the Bai and Perron (1998) test due to its widespread use in practice.

Following Pitarakis (2004), an alternative to HC inference is to use a feasible generalized least squares (FGLS) procedure for coefficient break inference. To our knowledge, study to date has not examined whether the use of FGLS improves forecast accuracy over OLS in the presence of possible structural breaks. Sect. 2.2 provides details of our FGLS structural break testing methodology. Denoting the period of a variance break as $t = T_v$ $(1 < T_v < T)$, it should be noted that if $\hat{T}_c \geq \hat{T}_v$, then the OLS and FGLS estimators will be identical in (3).

Clearly, even if a coefficient break has occurred, the point estimate $\hat{T}_c$ does not capture the uncertainty associated with break date estimation. To reflect this, we also investigate whether use of a confidence interval (or set) can improve forecast accuracy compared with a possibly poor single coefficient break date estimate. For convenience of exposition, assume that the dates within the confidence interval are contiguous and denote the interval as $[\hat{T}_{cL}, \hat{T}_{cU}]$, where $\hat{T}_{cL}$ and $\hat{T}_{cU}$ are the lower and upper bounds of the confidence interval, respectively (see Sect. 2.3). Treating each date in the interval as one of a sequence of choices for the potential break date, the corresponding post-break window forecasts can be averaged to yield the confidence interval forecast

$$\hat{y}_{T+1,CI} = \frac{1}{\hat{T}_{cU} - \hat{T}_{cL} + 1} \sum_{t=\hat{T}_{cL}+1}^{\hat{T}_{cU}+1} \hat{\beta}'_{t:T}\, \mathbf{x}_T. \tag{4}$$

The expression in (4) is also appropriately amended in the obvious way when the forecast is obtained by averaging over a (non-contiguous) confidence set for the break date.

Note that averaging as in (4) effectively gives greatest weight to sample observations for $t > \hat{T}_{cL}$, since these contribute to each forecast in the average, with progressively less weight given to observations earlier in the confidence interval or set. Since the interval or set will be longer when σ_t in (1) is larger or the magnitude of the break is smaller, these circumstances lead to the forecast in (4) giving relatively greater weight to sample observations for $t > \hat{T}_{cL}$ compared with a low volatility or large break setting. In other words, in circumstances when the timing of the break is doubtful, greater weight is placed on observations that can be reliably classified as post-break, but with some weight also placed on earlier observations that also fall within the interval or set.

To shed light on the nature of a detected coefficient break and to gain (potential) efficiency in estimation when the null hypothesis of no break is rejected, we also examine whether testing for change in the individual coefficients of the model improves forecast accuracy. To allow for possible variance change, a standard HC t test is applied to each individual coefficient in the model to examine whether the values differ in the pre- and post-break sub-samples, treating the coefficient break date as known. If all changes are significant, forecasts are obtained using the post-break sub-sample, as in (3). Otherwise, the coefficient with the least significant change is restricted to be constant over time and the model is re-estimated. The remaining coefficients are again tested individually and the procedure continues until all remaining coefficients are either specified as constant or exhibit significant change at the estimated break date. If the model reduces to one in which only one coefficient has a break and the change in

this coefficient is not significant, the whole sample forecast is used despite the initial finding of a coefficient break. This stepwise testing approach is combined with the confidence interval forecast of (4) by conducting stepwise coefficient equality testing at each potential break date in the interval.

When variance breaks are explicitly taken into account through FGLS estimation, the procedure just described is applied using standard (OLS) t tests for breaks in each individual coefficient of the FGLS-transformed model. It might also be noted that the application of individual coefficient tests implies the use of some pre-break data in obtaining $\hat{\beta}_{\hat{T}_c+1:T}$ and in this case the OLS and FGLS estimators are no longer necessarily identical when $\hat{T}_c \geq \hat{T}_v$.

The methodology employed for structural break testing is discussed in the next subsection, with the subsequent subsection considering the construction of confidence intervals for the break date. Throughout the paper, all hypothesis tests are conducted at the nominal (asymptotic) 5% level of significance[4] and the nominal confidence of all confidence intervals/sets is 95%. To ensure sufficient observations are available for estimation and structural break inference, the range of possible break dates is restricted to $\underline{w} < \hat{T}_i < T - \underline{w}$ for both the coefficients and variance ($i = c$ or v). Our results use $\underline{w} = 0.1T$, so that the minimum estimation window for the post-sample estimator of (3) is 10% of the full sample data. When a confidence interval forecast is employed, $\hat{T}_{cU} + 1$ in the summation of (4) is replaced by $\min(\hat{T}_{cU} + 1, T - \underline{w})$ and if $T - \underline{w} < \hat{T}_{cU} + 1$ the denominator is correspondingly adjusted.

Through a simulation analysis in Sect. 4, the performance of the methods we propose are compared with trade-off and cross-validation procedures proposed in an influential paper by Pesaran and Timmermann (2007) . Our comparison also includes the forecast combination method that averages over all possible estimation windows, which is proposed by Pesaran and Timmermann (2007) and analytically developed by Pesaran and Pick (2011).[5] Further information relating to these methods is provided in "Appendix A", but note that the cross-validation methods we employ do not use estimated break date information.[6]

2.2 Structural break testing methodology

The most commonly employed methodology for structural break inference in econometrics is that of Bai and Perron (1998), and our approach is based on their methodology. We investigate both the HC approach of Bai and Perron (1998) and also a two-step FGLS procedure when testing for a coefficient break in (1). In both cases, the outcome of the test determines whether the full-sample or post-break forecast is employed.

[4] Preliminary investigation indicated that the forecast performance was little affected by the adoption of a 10% significance level.

[5] For the case of multiple breaks, Tian and Anderson (2014) find that weighting forecasts according to the location of the window within the sample performs well. However, we do not include this as our interest focuses on single breaks.

[6] Our initial investigations also included the form of cross-validation that uses estimated break date information, with results qualitatively similar to those shown.

Our two-step FGLS method is based on Bataa et al. (2013), who generalize an approach suggested by Pitarakis (2004) to allow the possibility that coefficient and variance breaks are not necessarily concurrent.[7] In outline, our procedure is:

Step 1 *Preliminary coefficient break test* The Bai and Perron (1998) structural break testing procedure is applied to β of (1) employing HC inference. After allowing for a detected coefficient break, the residuals ($\hat{\varepsilon}_t$) are employed in the test regression

$$\sqrt{\frac{\pi}{2}}|\hat{\varepsilon}_t| = \zeta + \epsilon_t \tag{5}$$

to which the homoskedastic testing is applied. If a break is detected in (5) at $\hat{T}_v$, the estimates of ζ from the regimes $t = 1, \ldots, \hat{T}_v$ and $\hat{T}_{v+1}, \ldots, T$ yield the estimated standard deviations for the respective detected variance regimes.

Step 2 *Re-assessment of coefficient break* If a break is detected for (5), the FGLS transformation is applied to the data; otherwise the original data are used. The presence of a coefficient break is then re-assessed employing homoskedastic inference.

The absolute value of the residuals from the initial OLS estimation is used in (5) rather than the mean of squared residuals because this is more robust to non-normality (Davidian and Carroll 1987; McConnell and Perez-Quiros 2000) and, further, our preliminary analysis found it yielded a better variance break date estimate.

When testing for a coefficient break using either HC inference or in the two-step procedure, the distribution of regressors is allowed to change at the break date,[8] but the disturbances are assumed to be serially uncorrelated. Since our interest focuses on the possibility of a single structural break in each of the coefficients and variance, a maximum of one break is considered in each step.

2.3 Confidence interval/set estimation

As already noted in the Introduction, after testing for coefficient breaks as described in Sect. 2.2, we employ two procedures for computing coefficient break dates confidence intervals/sets. The first is that of Bai (1997), which is widely available as part of the inference procedure of Bai and Perron (1998). The confidence interval is constructed using the asymptotic framework of break date estimation and relates to dates in the "neighbourhood" of the estimate $\hat{T}_c$ (see Bai 1997; Bai and Perron 1998); consequently, the confidence interval is contiguous around $\hat{T}_c$.

Despite the popularity of the Bai and Perron (1998) procedure, the coverage rates for the associated confidence intervals are often substantially below the nominal rates, as shown by Elliott and Müller (2007) and the simulation results of Bai and Perron

[7] Although Bataa et al. (2013) iterate between coefficient and variance breaks, we employ a two-step procedure because their results (Table 1) indicate that iteration has relatively little impact on the detection of variance breaks.

[8] Since lagged y_t is included as a regressor, a coefficient break in (1) implies a break in the regressor matrix one period later. Although the dates do not quite coincide, it is appropriate to allow for the distribution of regressors to change when testing for a coefficient break.

(2006). The approach of Elliott and Müller (2007) and Eo and Morley (2015) is to invert the test statistic for a break, yielding a confidence set (not necessarily an interval) for the break date. We employ the confidence set of Eo and Morley (2015), which employs the likelihood ratio test statistic, as they find it provides good coverage with a smaller set of potential break dates than that of Elliott and Müller (2007).

Whether two-step or HC inference is employed for coefficient break date estimation, the original data are used for computing the confidence interval or set, with some account taken of a possible variance break (albeit contemporaneous with any coefficient break) by computing the confidence interval or set allowing variances to change with coefficient break regimes. In both cases, the distribution of regressors is allowed to change at the break date.

3 Monte Carlo simulations

Monte Carlo simulations are conducted to evaluate the forecast methodologies proposed in Sect. 2, with these based on the simulation setup of Pesaran and Timmermann (2007), with similar settings also adopted by Clark and McCracken (2005) and Tian and Anderson (2014). However, in addition to the simulation settings in these papers, we also consider data generating processes (DGPs) that exhibit change only in the intercept and DGPs with changes in both coefficients and variances with these changes not necessarily occurring at the same time.

The basic DGP is the bivariate VAR(1) process

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha_{yt} \\ \alpha_{xt} \end{pmatrix} + \mathbf{A}_t \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{yt} \\ u_{xt} \end{pmatrix} \tag{6}$$

with coefficient matrix

$$\mathbf{A}_t = \begin{pmatrix} \beta_{11t} & \beta_{12t} \\ 0 & \beta_{22t} \end{pmatrix} \tag{7}$$

so that x Granger causes y and not vice versa, and disturbances that are normally distributed with covariance matrix

$$\Sigma_t = E\left[\begin{pmatrix} u_{yt} \\ u_{xt} \end{pmatrix} \begin{pmatrix} u_{yt} \\ u_{xt} \end{pmatrix}' \right] = \begin{pmatrix} \sigma_{yt}^2 & 0 \\ 0 & 1 \end{pmatrix}. \tag{8}$$

If $\mathbf{A}_t$ is time invariant and has eigenvalues strictly less than unity, the unconditional mean vector corresponding to (6) and (7) is

$$\mu_t = (I - A_t)^{-1}\alpha_t = (I - A_t)^{-1}(\alpha_{yt}\ \alpha_{xt})'. \tag{9}$$

However, change in either α_t or A_t leads to change in the corresponding steady-state means of (9).

Our interest is in situations where the marginal distribution for x_t (the driving variable) is constant over time, but that for y_t can exhibit structural breaks in the

coefficients and/or disturbance variance. The cases considered (where the DGPs refer to Table 1) are:

1. All coefficients constant over time, with $\alpha = (0.5\ 0.5)'$, $\beta_{11} = 0.9$, $\beta_{12} = 1$, $\beta_{22} = 0.9$, together with time-invariant $\sigma_y = 1$ (DGP1), increasing variance (DGP2) or decreasing variance (DGP3);
2. Dynamics that exhibit change (through β_{11t} and/or β_{12t}) but μ_t constant over time with values as in #1 (hence α_t changes), with either constant or changing variance (DGP4 to DGP10);
3. Means that exhibit change through a break in the value of α_y but constant dynamics as in #1, with either constant or changing variance (DGP11 to DGP14).

Details of parameter values are provided in Table 1.

The benchmark case, which uses OLS and all observations, is anticipated to provide the best forecasting performance for DGP1. DGPs 2 and 3 illustrate the effects of increasing and decreasing disturbance variances, with the post-break standard deviations increasing by a factor of 4 and declining by a factor of two, respectively. In DGPs 4 and 5, the autoregressive coefficient, β_{11t}, decreases, with these labelled as small and large changes after the break, respectively. Similarly, the effects of small and large increases in the coefficient of the lagged exogenous variable, β_{12t}, are considered in DGPs 6 and 7. DGP8 combines DGP4 and DGP7, with a single break affecting both β_{11t} and β_{12t} coefficients simultaneously. DGP9 and DGP10 introduce changes in both coefficients and variances, by combining DGP8 with 2 and 3, respectively. A break affecting only the mean of the series, namely an intercept shift, is illustrated in DGPs 11 and 12, considered as small and large, respectively. Finally, DGPs 13 and 14 examine situations in which a large mean increase is combined with increasing or decreasing variances after the break.

In order to examine the sensitivity of the break point location for forecasting performance, the simulations consider a single coefficient break occurring at $0.25T$, $0.5T$ or $0.75T$ of the full sample of T observations. Further, in DGPs with changing variances, the single variance break occurs in the middle of the sample in combination with each coefficient break location or at $0.75T$ in combination with a mid-point coefficient break. Therefore, we consider scenarios where the coefficient and variance breaks occur either concurrently or at different times, and also whether a variance break precedes or follows a coefficient break. It can be noted that breaks occur only in the equation for y_t in (6) and inference is applied to this equation only.

We employ the Bai and Perron (1998) procedure to test for breaks in which we allow the possibility of one break with trimming $\epsilon = 0.10$ (10% of the full sample). All hypothesis tests are conducted at a nominal 5% level of significance, using the asymptotic critical values provided by Bai and Perron (1998), while the confidence intervals/sets have nominal 95% coverage. In line with Pesaran and Timmermann (2007), the methods of cross-validation or averaging across all samples to T consider a minimum sample size of $0.1T$, while cross-validation reserves $0.25T$ observations for an out-of-sample evaluation.[9]

[9] Although the results presented in the paper make the realistic assumption that neither the presence nor the date of any break is known, we also obtained results for known break dates. Footnotes below sometimes refer to these results, which can be obtained from the authors on request.

Table 1 Details of data generating processes

	Parameters	α_{yt}		β_{11}		β_{12}		σ_{yt}	
DGPs	Comments/regimes	R1	R2	R1	R2	R1	R2	R1	R2
DGP1	No break	0.5	0.5	0.9	0.9	1	1	1	1
DGP2	Increase in σ_{yt}	0.5	0.5	0.9	0.9	1	1	1	4
DGP3	Decrease in σ_{yt}	0.5	0.5	0.9	0.9	1	1	1	0.5
DGP4	Small break in β_{11}	0.5	11.5	0.9	0.7	1	1	1	1
DGP5	Large break in β_{11}	0.5	22.5	0.9	0.5	1	1	1	1
DGP6	Small break in β_{12}	0.5	−2	0.9	0.9	1	1.5	1	1
DGP7	Large break in β_{12}	0.5	−4.5	0.9	0.9	1	2	1	1
DGP8	Break in β_{11} and β_{12}	0.5	6.5	0.9	0.7	1	2	1	1
DGP9	Break in β_{11}, β_{12}+increase in σ_{yt}	0.5	6.5	0.9	0.7	1	2	1	4
DGP10	Break in β_{11}, β_{12}+decrease in σ_{yt}	0.5	6.5	0.9	0.7	1	2	1	0.5
DGP11	Small mean break in α_{yt}	0.5	1.3	0.9	0.9	1	1	1	1
DGP12	Large mean break in α_{yt}	0.5	2.5	0.9	0.9	1	1	1	1
DGP13	Large mean break in α_{yt}+Increase in σ_{yt}	0.5	2.5	0.9	0.9	1	1	1	4
DGP14	Large mean break in α_{yt}+Decrease in σ_{yt}	0.5	2.5	0.9	0.9	1	1	1	0.5

The values of coefficients and disturbance standard deviation before and after the break are given in R1 and R2 columns, respectively

To assess the impact of the sample size on forecasting performance, $T = 100$ and 200 are employed in the simulations. The DGP process in Eq. (6) starts from its pre-break unconditional mean. Specifically, each replication of each DGP is initialized using $(y_{t-1}\ x_{t-1})' = \alpha_t = (0.5\ 0.5)'$, and simulating $T_0 + T + 3$ observations for the corresponding DGP with $T_0 = 100$. After discarding the first T_0 observations, the observations $1, \ldots, T$ are used for the parameter estimation to generate the forecasts.[10] In all cases, 5000 replications are employed.[11]

Forecast accuracy is assessed using the empirical Mean Squared Forecast Error (MSFE), namely the average squared difference between forecast and realized values, computed as

$$\text{MSFE} = 1/S \sum_{i=1}^{S} (y_{T+1} - \hat{y}_{T+1})^2 \tag{10}$$

where S denotes the number of simulations. In the results reported, the computed MSFE for each method is divided by the MSFE of the benchmark model which ignores the presence of possible breaks by applying the full sample OLS estimator. Ratios lower than 1 indicate better performances of the corresponding methods than the benchmark, and higher than 1 points to worse performances compared to the benchmark model.

4 Simulation results

Our simulation results are discussed in the first subsection for the special case where coefficient and disturbance breaks, when they occur, are concurrent at the mid-point of a sample of $T = 100$ observations. Subsequent subsections consider non-concurrent breaks for one-step ahead forecasts and $T = 100$ and (finally) larger sample results for $T = 200$.

4.1 Concurrent mid-sample breaks

Table 2 provides background inference results, while Table 3 and Appendix Table 8 report relative MSFEs (in relation to the benchmark model) for a range of forecast methods when any (coefficient or variance) break occurs in the middle of the sample of $T = 100$ observations. Table 3 and Appendix Table 8 differ only in that the latter sets the forecast period disturbance $\varepsilon_{T+1} = 0$ in (1).[12] By focusing on the trade-off between bias in coefficient estimation when pre-break information is included and efficiency gains from these additional observations (Pesaran and Timmermann 2007), Appendix

[10] We also computed forecasts $\widehat{y}_{T+h}$ ($h = 2, 3$) by estimating the true (constant parameter) AR(1) model for x_t and using this in conjunction with the specification of the y_t model resulting from the procedures considered in Sect. 2 to obtain iterated multi-horizon forecasts. The MSFE results show the same patterns as those for one-step ahead and are available from the authors on request.

[11] The initial seed is set for each DGP so that all forecasting methods are evaluated based on exactly the same sample data.

[12] We thank a referee for this suggestion.

Table 8 provides a clearer distinction between methods. However, the apparent gains it indicates are unattainable in realistic settings and hence our discussion focuses on Table 3.

From the empirical coefficient structural break test rejection rates (the percentage of cases in which a break is detected) in Table 2, it is evident that the two-step procedure improves on HC inference by substantially reducing the number of over-rejections in the DGPs with constant coefficients (DGP1-DGP3) while increasing the number of rejections in the DGPs where the HC test has low empirical power in DGPs 9 and 13. Since individual coefficients are tested only when an overall break is detected, these rates (expressed as the percentage of total replications for which constancy is rejected) are always less than the overall rejection rate. The empirical variance test rejection rate is also shown for the two-step method of Sect. 2.2.

For convenience, we use the abbreviation CI in Table 2 to refer to both confidence intervals and sets, with those associated with Bai and Perron (1998) and Eo and Morley (2015) referred to as BP and EM, respectively. In line with previous studies (including Bai and Perron 2006; Elliott and Müller 2007; Chang and Perron 2018; Bai 1997; Eo and Morley 2015), the EM set almost always has greater coverage of the true break date than the nominal 95%, whereas the BP interval exhibits under-coverage and this is often substantial. Although used only when coefficient constancy is rejected, the average CI length is shown for cases where coefficient breaks are rejected and where they are not, together with all cases.[13] With no coefficient or variance break in DGP1 and allowing for 10% trimming, the vast majority of sample points admissible as potential breaks fall within the EM set; hence it performs well in indicating the lack of information in the data about a coefficient break, whereas the BP interval includes substantially fewer observations; these results are in line with Eo and Morley (2015) and carry over in the presence of a variance break (DGPs 2, 3). Across all DGPs and methods, inclusion of only cases where the coefficient test is rejected reduces (or leaves unchanged) the average CI lengths, because these are cases where stronger evidence of a break is detected. As anticipated, both methods include more potential break points in the interval/set when the break is small compared with large breaks.

Due to the over-sizing of the HC break test (Table 2), post-break OLS estimation leads to poor forecasting results for DGP1 (no break) and DGP2 (variance increase) in Panel A of Table 3 compared with the full-sample benchmark model. With improved inference on coefficient breaks, the two-step method leads to improved accuracy in Panel B for these DGPs. The effect of over-sizing on forecast accuracy is less severe in DGP3 since the variance decrease implies that the less volatile sub-sample is employed for estimation when a coefficient break is erroneously detected. In line with Pesaran and Timmermann (2007),[14] the trade-off method, which also hinges on the estimated break date, improves accuracy over the simple use of the post-break sample, but averaging based on either the BP interval or (particularly) the EM set does better, with further improvement when combined with stepwise coefficient testing. Indeed, averaging over the EM set combined with stepwise coefficient testing reduces the MSFE to a little

[13] We thank a referee for prompting us to this discussion and additional results.

[14] Note, however, that Pesaran and Timmermann (2007) examine only concurrent coefficient and variance breaks and hence do not consider HC or two-step inference.

Table 2 Inferences on concurrent mid-sample breaks, $T = 100$

	Coefficients constant			Overall mean constant							Dynamics Constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
HC coefft. test rejections (%)														
Overall test	18.1	27.8	14.5	98.6	100.0	89.4	100.0	100.0	58.1	100.0	62.0	98.6	58.5	99.9
Intercept	15.4	23.7	12.5	98.3	100.0	77.0	98.6	95.4	48.5	96.4	45.5	78.7	50.7	82.1
β_{11}	14.9	24.6	11.7	97.7	99.9	25.1	13.6	99.8	57.3	99.8	25.9	21.0	40.6	18.6
β_{12}	10.4	15.8	8.6	13.5	12.0	84.3	100.0	99.9	50.6	99.9	17.8	17.7	21.0	17.1
Two-step test rejections (%)														
Overall test	8.3	16.5	9.3	98.5	100.0	88.5	100.0	100.0	73.5	100.0	58.4	98.5	69.2	99.9
Intercept	6.9	13.4	7.7	98.1	100.0	76.1	98.6	95.4	60.0	96.7	42.8	78.7	58.6	82.5
β_{11}	6.4	13.5	6.7	97.6	99.9	23.6	12.6	99.8	72.7	99.9	22.0	20.3	39.7	19.2
β_{12}	4.2	8.8	4.9	12.3	10.6	83.2	100.0	99.9	66.8	100.0	14.9	16.2	20.0	18.1
Variance break rejections (%)	6.0	100.0	91.6	5.4	5.3	5.7	5.8	5.6	100.0	94.5	5.9	5.6	100.0	94.7
CI coverage (% of test rejections)														
BP interval: HC	NA	NA	NA	81.4	89.3	71.1	82.0	86.7	68.8	86.2	65.4	92.9	50.2	96.1
EM interval: HC	NA	NA	NA	97.4	98.6	95.5	98.6	98.2	99.9	100.0	90.6	98.2	100.0	100.0
BP interval: two-step	NA	NA	NA	81.5	89.4	71.9	82.0	86.7	73.9	86.5	69.5	92.9	68.7	96.6
EM interval: two-step	NA	NA	NA	97.4	98.6	95.5	98.6	98.2	99.7	100.0	90.6	98.2	99.9	100.0

Table 2 continued

	Coefficients constant			Overall mean constant							Dynamics Constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
CI length (average, test rejections)														
BP interval: HC	12.6	13.6	13.0	9.8	5.8	16.2	7.1	8.7	21.7	5.4	17.7	7.1	16.1	5.2
EM interval: HC	49.3	61.3	55.3	11.7	5.1	24.7	9.5	9.0	33.4	11.2	32.3	8.1	41.6	9.9
BP interval: two-step	15.4	18.1	23.4	9.8	5.8	16.3	7.1	8.7	27.7	5.3	18.3	7.1	21.4	5.2
EM interval: two-step	34.3	41.0	43.3	11.6	5.1	24.3	9.5	9.0	22.8	11.1	30.2	8.1	27.3	9.8
CI length (average, test non-rejections)														
BP interval: HC	36.7	34.9	41.2	34.4	NA	34.3	NA	NA	37.6	NA	31.1	15.4	29.7	17.0
EM interval: HC	68.5	63.9	64.0	52.0	NA	57.0	NA	NA	46.8	NA	60.0	46.3	56.5	54.6
BP interval: two-step	36.1	35.3	39.9	33.1	NA	33.1	NA	NA	37.9	NA	30.2	15.9	30.7	14.0
EM interval: two-step	68.1	50.8	60.3	54.7	NA	57.3	NA	NA	34.2	NA	61.0	46.6	37.7	50.7
CI length (average, all cases)														
BP interval: HC	32.3	29.0	37.1	10.1	5.8	18.1	7.1	8.7	28.4	5.4	22.8	7.2	21.7	5.2
EM interval: HC	65.0	63.2	62.8	12.2	5.1	28.1	9.5	9.0	39.0	11.2	42.8	8.6	47.8	9.9
BP interval: two-step	34.4	32.5	38.4	10.2	5.8	18.2	7.1	8.7	30.4	5.3	23.2	7.2	24.3	5.2
EM interval: two-step	65.3	49.2	58.8	12.2	5.1	28.1	9.5	9.0	25.8	11.1	43.0	8.6	30.5	9.8

The details of the DGPs are presented in Table 1. T is the total sample size and the true coefficient and disturbance breaks (if they occur) are in the middle of the sample. NA refers to not applicable as the coverage of confidence interval for a coefficient break is not defined in the DGPs with no coefficient break. NA is also used for the average length of the confidence interval in non-rejection cases when the coefficient break test rejects constancy in all simulations. BP interval refers to confidence interval estimated by Bai and Perron (1998) and EM interval refers to confidence set computed by Eo and Morley (2015) procedure.

Table 3 MSFE ratios: concurrent mid-sample breaks, $T = 100$

	Coefficients constant			Overall mean constant							Dynamics constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
Panel A: HC break inference														
Post-break	1.124	1.185	1.019	0.610	0.455	0.812	0.446	0.508	1.130	0.201	1.008	0.673	1.200	0.322
Trade-off	1.075	1.122	1.007	0.612	0.477	0.804	0.450	0.514	1.075	0.214	0.977	0.672	1.121	0.330
Stepwise testing	1.113	1.177	1.018	0.606	0.451	0.818	0.445	0.510	1.133	0.202	0.984	0.666	1.180	0.331
BP confidence interval average	1.070	1.098	1.009	0.617	0.470	0.806	0.449	0.517	1.046	0.202	0.969	0.670	1.081	0.322
EM confidence interval average	1.029	1.021	1.002	0.607	0.455	0.796	0.445	0.508	0.998	0.202	0.964	0.677	1.013	0.322
BP confidence interval and stepwise testing	1.064	1.092	1.007	0.611	0.458	0.810	0.447	0.515	1.055	0.202	0.954	0.663	1.071	0.327
EM confidence interval and stepwise testing	1.024	1.015	1.000	0.606	0.453	0.796	0.445	0.510	1.004	0.202	0.950	0.666	1.005	0.327
Panel B: Two-step break inference														
Post-break	1.054	1.082	1.021	0.610	0.455	0.812	0.447	0.509	1.096	0.201	0.990	0.674	1.139	0.319
Trade-off	1.033	1.055	1.022	0.614	0.478	0.805	0.451	0.514	1.033	0.210	0.969	0.673	1.065	0.325
Stepwise testing	1.052	1.085	1.034	0.606	0.450	0.818	0.446	0.509	1.084	0.201	0.975	0.666	1.098	0.320

Table 3 continued

	Coefficients constant			Overall mean constant							Dynamics constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
BP confidence interval average	1.032	1.045	1.026	0.618	0.470	0.806	0.449	0.517	1.029	0.202	0.963	0.671	1.044	0.320
EM confidence interval average	1.021	1.017	1.005	0.609	0.455	0.797	0.445	0.508	0.996	0.202	0.963	0.678	1.016	0.322
BP confidence interval and stepwise testing	1.032	1.047	1.024	0.610	0.456	0.809	0.448	0.516	1.031	0.202	0.949	0.663	1.032	0.324
EM confidence interval and stepwise testing	1.019	1.014	1.001	0.605	0.453	0.796	0.446	0.510	1.002	0.202	0.946	0.666	0.998	0.325
Panel C: Unknown break inference														
Cross-validation (CV)	1.036	1.043	1.009	0.624	0.471	0.800	0.460	0.525	1.005	0.207	0.952	0.698	1.033	0.333
Average over all windows (CV weight)	1.013	1.024	0.981	0.685	0.572	0.808	0.529	0.590	0.965	0.270	0.921	0.737	0.992	0.417
Average over all windows (equal weight)	1.039	1.051	0.997	0.664	0.555	0.804	0.527	0.578	0.980	0.319	0.927	0.720	1.009	0.430

The details of the DGPs are presented in Table 1. T is the total sample size and the true break date is in the middle of the sample. MSFEs are reported relative to the associated MSFE based on the full sample benchmark model

over one in both panels, implying a relatively small accuracy loss relative to using the full-sample OLS estimator. The methods in Panel C (which do not rely on an estimated break date, namely cross-validation and averaging over all windows) also perform relatively well for these DGPs: these are more accurate than simple post-break (OLS or FGLS) estimation, with the method that averages with cross-validation weights yielding the most accurate forecasts for DGPs 1 and 3 (though the former is less accurate than the benchmark model). The effects discussed here and elsewhere in this subsection are emphasized in Appendix Table 8, where the forecast period disturbance is shut down.

When a break increases the dynamic coefficients with a time-invariant disturbance variance (DGPs 4–8, Table 3), it is unsurprising that all forecasting methods which take account of a possible break perform better than the benchmark. However, it is worth noting that in these DGPs the presence of a coefficient break is relatively easy to detect even if it is "small" (DGPs 4 and 6, Table 2), since the variance is constant while the intercept changes in addition to the dynamic coefficients due to the constant mean assumption; see (9). Therefore, the forecast accuracy results for these DGPs in Table 3 are very similar across Panels A and B. Although cross-validation performs well for these DGPs, the EM set average is more accurate.

Almost all methods of Panels A and B are less accurate than the full-sample OLS estimator for DGP9. The variance break makes it more difficult to detect the coefficient break (compare the overall coefficient test rejections for DGP9 with DGP8 in Table 2) and, in any case, the benefits of detecting the coefficient break can be out-weighed by using only noisier post-break data.[15] Nevertheless, two-step inference reduces the MSFE in Panel B compared with HC inference in Panel A of Table 3, while using either the EM set average or the methods in Panel C results in forecasts that are very close in accuracy to (or for averaging over windows and using cross-validation weights, better than) the benchmark case. On the other hand when the variance decreases in DGP10, using break information is highly beneficial relative to use of full-sample OLS because the coefficient break is always detected and the latter part of the sample is less noisy. All methods that use break information yield very similar results and reduce MSFEs by about 80% compared with the benchmark, while averaging over windows or using cross-validation weights does less well.

Finally, when a coefficient break in DGPs 11–14 affects only the intercept, Table 2 shows that stepwise coefficient testing assists in pin-pointing the nature of the break, with constancy of the intercept rejected more frequently than constancy of either dynamic coefficient. However, when the variance declines (DGP14) stepwise testing can result in slightly increased MSFE values, whether applied alone or in combination with averaging over a confidence interval or set; this also applies in DGP9 where breaks occur in both lag coefficients and the variance increases. In general, small mean breaks are hard to identify accurately and ignoring rather than modelling them often leads to more accurate forecasts (Pesaran and Timmermann 2005; Boot and Pick 2020). In line with such findings, the smallest MSFE values for DGP11 in Table 3 are achieved

[15] Even with known break dates, the relative MSFE values for DGP9 are close to, and sometimes a little larger than unity; the relative MSFE for the post-break estimator is then 0.980. Similar comments apply for DGP13, where the post-break estimator with known break dates is 1.016.

by methods which average over all windows, followed by the stepwise testing method combined with averaging over the EM confidence set.

The results for DGPs 9 and 13 emphasize that, with either HC or two-step inference, use of a post-break estimator can lead to a deterioration in forecast accuracy compared with full-sample OLS when a coefficient break occurs. However, averaging over the EM set combined with testing down effectively eliminates this deterioration. In these cases averaging using cross-validation weights or over all windows (Panel C) also performs very well, with the latter having the lowest relative MSFE across all methods considered. Many of the forecasts which are averaged over all windows include a substantial number of less volatile pre-break observations and their associated smaller forecast errors help to reduce overall forecast errors. On the other hand, these two averaging methods of Panel C perform substantially worse than methods that explicitly use coefficient break date inference in both DGPs 10 and 14, when coefficients change alongside a decline in the disturbance variance.

The results just discussed shed new light on the importance of break inference for forecasting. In particular, the better inference properties of the EM confidence set compared with the BP confidence interval (Table 2) yield more accurate averaged forecasts in almost all DGPs in Table 3.

4.2 Non-concurrent breaks

The forecast accuracy results in Table 3 represent a special case in which, when both occur, the coefficient and disturbance variance breaks are concurrent. However, in practice such breaks may not coincide and the two-step structural break testing method of Sect. 2.2 is designed to cope with this. To assess the impacts of different locations of a coefficient break point, Table 4 examines cases where this occurs earlier (at $0.25T$) in the upper part of the table and later ($0.75T$) in the lower part, with any variance break applying at the sample mid-point; once again $T = 100$.[16] Results for DGPs 1–3 are excluded since these are unchanged from Table 3. Further, results for the large coefficient break cases of DGPs 5 and 7 are omitted, since the pattern of results carries over from Table 3. Although results are unchanged for DGPs 8 and 12, these are included to facilitate comparison with DGPs 9–10 and 13–14, respectively. To conserve space, methods that employ the BP interval are also excluded, as these are almost always inferior to those using the EM set, and stepwise coefficient testing is included only in combination with the EM set, as (in common with Table 3) this effectively dominates use of stepwise testing without averaging over possible break dates.

Despite coefficient and disturbance variance break dates not being concurrent, the results for the early coefficient break case in Panels A and B of Table 4 show broadly similar patterns to those in Table 3. In particular, two-step inference is beneficial when breaks are small or especially when the disturbance variance increases (DGPs 4, 6, 9, 13). However, unlike in Table 3, forecast accuracy also improves when using two-step over HC inference when the variance declines (DGPs 10 and 14). In circumstances

[16] Results corresponding to Table 4 and also Table 5 but setting $\varepsilon_{T+1} = 0$ (as in Appendix Table 8) are available from the authors on request.

when detection and dating of a coefficient break are difficult (DGPs 4, 6, 9, 11, 13), averaging over the EM confidence set reduces MSFE relative to using a point estimate of the break date and a post-break estimator. Nevertheless, even for DGPs 11–13, and in contrast to the corresponding cases in Table 3, there is generally little benefit here from using stepwise coefficient testing with the EM set, presumably because there is now a larger number of observations available after the true coefficient break date. Averaging using cross-validation weights performs well and yields the lowest MSFE across all methods when the breaks are small in DGPs 6 and 11, and also for the mean shift and variance increase case of DGP13.

When the true coefficient break date occurs relatively late in the sample (lower part of Table 4), the use of HC versus two-step inference has little impact on the MSFE values. As noted in Sect. 2, when $\hat{T}_c \geq \hat{T}_v$ and for a given coefficient break date estimate, the post-break estimator (3) uses OLS whether HC or two-step inference is employed. Since the true $T_c > T_v$ here, less gain may be anticipated from the two-step method compared to the upper part where $T_c < T_v$. Otherwise, the relative performances of the methods of Panels A and B are broadly similar to those for other coefficient break locations in Table 3 and the upper part of Table 4. However, the methods of Panel C are generally quite poor when the coefficient break occurs late in the sample. For the cross-validation methods, this is explained by the final 25% of the sample being reserved for a pseudo forecasting exercise, while averaging over all windows gives relatively less weight to the true post-break observations when the break occurs in the latter part of the sample.

Finally, Table 5 considers the case of a late $(0.75T)$ variance break in combination with a mid-point coefficient break; hence in common with the upper part of Table 4, the DGPs of Table 5 have $T_c < T_v$ when breaks in both components occur. Results are shown for the same methods as in Table 4, but for a different set of DGPs. In particular, DGPs 1, 4–8 and 11–12 have constant variance and hence have unchanged results from Table 3, and these results are not repeated. The results in Table 5 confirm the benefits of using two-step structural break over HC inference when the variance break (especially an increase) occurs after the coefficient break.[17] It is also noteworthy that averaging with cross-validation weights in Panel C also performs well when the disturbance variance increases, but less well in the presence of a variance decrease.

4.3 Larger sample size

The effects of a larger sample size on forecast performance are explored using $T = 200$ in Appendix Tables 9, 10 and 11, which show corresponding results to those of Tables 3, 4 and 5. Since break sizes are fixed across the sample sizes, it is not surprising that the larger sample improves estimation of the break date and hence there is relatively less gain from averaging over a confidence interval/set compared with using the post-break estimator. Nevertheless, gains typically apply when the EM confidence set is used and either there is no coefficient break (DGPs 1–3) or the variance increases alongside a coefficient break (DGPs 9 and 13), with very little or no loss of accuracy

[17] A similar pattern of results applies when the break dates are known, with the use of all two-step (FGLS) methods improving on the use of HC inference.

Table 4 MSFE ratios: non-concurrent breaks, $T = 100$

Forecast methods/DGPs	Overall mean constant					Dynamics constant			
	Small Break in β_{11}	Small Break in β_{12}	Break in β_{11} β_{12}	DGP 8 $\sigma \times 4$	DGP 8 $\sigma \times 0.5$	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 4	DGP 6	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
Tc = 0.25T, Tv = 0.50T, T = 100									
Panel A: HC break inference									
Post-break	0.724	0.924	0.702	1.144	0.380	1.041	0.790	1.188	0.482
Trade-off	0.728	0.911	0.704	1.087	0.385	1.010	0.788	1.117	0.484
EM interval average	0.719	0.904	0.702	1.006	0.372	0.988	0.792	1.013	0.475
EM interval and stepwise testing	0.719	0.905	0.704	1.007	0.376	0.987	0.789	1.011	0.483
Panel B: Two-step break inference									
Post-break	0.721	0.919	0.703	0.997	0.371	1.018	0.791	1.037	0.473
Trade-off	0.726	0.909	0.704	0.986	0.372	1.002	0.790	1.014	0.474
EM interval average	0.718	0.906	0.702	0.992	0.370	0.994	0.793	1.008	0.474
EM interval and stepwise testing	0.719	0.906	0.703	1.000	0.370	0.989	0.791	1.005	0.476
Panel C: Unknown break inference									
Cross-validation (CV)	0.739	0.915	0.729	1.015	0.381	0.990	0.808	1.031	0.483
Average over all windows (CV weight)	0.726	0.894	0.710	0.993	0.376	0.958	0.792	1.002	0.480
Average over all windows (equal weight)	0.737	0.913	0.722	1.016	0.388	0.981	0.807	1.028	0.489
Tc = 0.75T, Tv = 0.50T, T = 100									
Panel A: HC break inference									
Post-break	0.616	0.829	0.456	1.117	0.157	1.057	0.578	1.191	0.225
Trade-off	0.607	0.795	0.459	1.058	0.172	1.001	0.569	1.112	0.232

Table 4 continued

Forecast methods/DGPs	Overall mean constant					Dynamics constant			
	Small Break in β_{11}	Small Break in β_{12}	Break in β_{11} β_{12}	DGP 8 $\sigma \times 4$	DGP 8 $\sigma \times 0.5$	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 4	DGP 6	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
EM interval average	0.602	0.785	0.452	0.990	0.162	0.974	0.591	1.007	0.254
EM interval and stepwise testing	0.593	0.788	0.462	0.994	0.163	0.945	0.557	0.998	0.242
Panel B: Two-step break inference									
Post-break	0.616	0.838	0.458	1.109	0.157	1.044	0.578	1.185	0.224
Trade-off	0.609	0.807	0.461	1.049	0.173	0.997	0.569	1.097	0.232
EM interval average	0.605	0.798	0.454	0.997	0.162	0.978	0.592	1.012	0.254
EM interval and stepwise testing	0.595	0.798	0.466	1.000	0.164	0.946	0.554	1.001	0.243
Panel C: Unknown break inference									
Cross-validation (CV)	0.812	0.805	0.670	1.005	0.473	0.944	0.729	1.023	0.523
Average over all windows (CV weight)	0.870	0.851	0.812	0.984	0.715	0.928	0.770	0.998	0.604
Average over all windows (equal weight)	0.736	0.777	0.649	0.968	0.494	0.900	0.674	1.000	0.462

The details of the DGPs are presented in Table 1. T is the total sample size, T_c is the location of true coefficient break and T_v is the location of true disturbance variance break

Table 5 MSFE ratios: late variance breaks, $T = 100$

Forecast methods/GDPs	Coefficients constant		Overall mean constant		Dynamics constant	
	σ x 4	σ x 0.5	DGP 8 σ x 4	DGP 8 σ x 0.5	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 2	DGP 3	DGP 9	DGP 10	DGP 13	DGP 14
Tc = 0.50T, Tv = 0.75T, T = 100						
Panel A: HC break inference						
Post-break	1.195	1.036	1.150	0.220	1.197	0.349
Trade-off	1.121	1.016	1.080	0.228	1.108	0.349
EM confidence interval average	1.035	1.010	0.995	0.218	1.019	0.347
EM confidence interval & stepwise testing	1.027	1.004	1.000	0.219	1.011	0.348
Panel B: Two-step break inference						
Post-break	1.099	1.009	0.955	0.216	1.008	0.341
Trade-off	1.060	1.010	0.941	0.221	0.990	0.341
EM confidence interval average	1.040	1.007	0.950	0.216	1.001	0.343
EM confidence interval and stepwise testing	1.030	1.002	0.955	0.217	0.985	0.344
Panel C: Unknown break inference						
Cross-validation (CV)	1.030	1.013	0.978	0.220	1.025	0.355
Average over all windows (CV weight)	1.019	0.997	0.954	0.289	0.989	0.436
Average over all windows (equal weight)	1.050	1.001	0.970	0.321	1.010	0.431

The details of the DGPs are presented in Table 1. T is the total sample size, T_c is the location of true coefficient break and T_v is the location of true disturbance variance break

in other cases. It may also be noted that, for this larger sample size and a relatively parsimonious model, the performance of stepwise testing (either alone or in combination with averaging over a confidence interval or set) performs well only for pure intercept shifts when the variance is constant or increasing (DGPs 11–13).

It is also not surprising that the two-step method achieves less gain in forecast accuracy over the use of HC coefficient break inference for this larger sample size, with forecast gains particularly apparent for methods that rely on a point estimate of the coefficient break date (that is, the post-break, trade-off and stepwise coefficient testing methods) and when there is no parameter break of either type (DGP1) or the disturbance variance increases (DGPs 9 and 13). The only exception to this statement is for DGP9 in Appendix Table 10, where a coefficient break at $0.75T$ is combined with a variance break at the sample mid-point; in this case HC inference leads to better forecast performance. On the other hand, when the timing of the two breaks is reversed in Appendix Table 11 , two-step inference leads to substantially more accurate forecasts for this DGP.

Finally, considering the methods of Panel C that do not use a break date estimator, while the cross-validation methods perform relatively well in relation to other methods in some cases, use of averaging over the EM confidence provides more accurate forecasts overall.

4.4 Summary

The main findings from the simulations are as follows:

1. Although our context of possibly distinct coefficient and variance breaks differs from other studies (Pesaran and Timmermann 2004, 2007; Tian and Anderson 2014) our results underline their finding that forecast accuracy gains can be achieved over the use of a simple post-break coefficient estimator.
2. Averaging across potential break dates as in (4) typically improves forecast accuracy relative to methods based on a point estimate of the break date, including trade-off and stepwise coefficient testing methods. For this purpose, the EM confidence set performs better overall than the BP confidence interval, due essentially to the poorer coverage of the true break date (when there is one) and smaller length of the latter compared with the former.
3. Employing two-step rather than HC inference for detecting and dating a coefficient structural break generally reduces forecast errors in moderate or small samples when the coefficients are, in fact, constant over time or when the disturbance variance exhibits change (particularly an increase) at the same time as or at a period subsequent to any coefficient break. Further, even when it does not improve forecast accuracy, two-step inference leads to little or no accuracy deterioration, because two-step inference reduces over-rejections when no coefficient break occurs and also more often detects true breaks when the variance increases during the sample period.
4. When the disturbance variance is constant or increases over the sample period, attempting to pinpoint the nature of a coefficient break by testing down typically leads to improved forecast accuracy over treating all coefficients as changing.

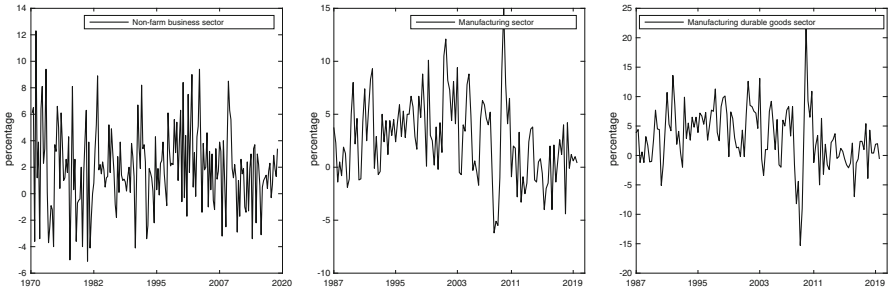

Fig. 1 US productivity growth series

However, when either HC or two-step inference is used, benefits do not reliably accrue when the variance decreases, irrespective of whether this is adopted in conjunction with interval/set averaging.

5. Averaging using cross-validation weights (recommended by Pesaran and Timmermann (2007)) performs well relative to other methods in the presence of a disturbance variance increase, but relatively poorly when the variance decreases and is not suited to cases where a coefficient break occurs late in the sample period.

In summary, combining information from structural break tests and confidence intervals/sets can improve forecast accuracy, particularly in small samples.

5 Application to US productivity growth

In order to investigate how well our proposed methods work with observed data, we undertake a forecasting exercise for US labour productivity growth. The apparent slowdown of US productivity growth in the current century and its possibly changing dynamics are well documented in a number of studies [and more] (Syverson 2017; Jorgenson et al. 2008; Benati 2007; Hansen 2001), so it is of interest to see how well the range of methods we consider in Sect. 4 perform in a pseudo forecasting exercise for such series.

We analyse three measures of labour productivity growth (growth in real output per hour worked) published by the Bureau of Labor Statistics,[18] namely productivity in the non-farm business sector, manufacturing sector and manufacturing durable goods sector. Data are quarterly seasonally adjusted values for the percentage change at an annual rate. The non-farm business sector series covers the period between 1970Q1 and 2018Q4, with the other two series available from 1987Q1 to 2018Q4; see Fig. 1. The final 20 quarterly observations are used for evaluating out-of-sample forecasting performance. Although the selection of 20 quarters reflects the total sample sizes available, especially for the sectoral series, a robustness analysis is discussed below in relation to this choice.

Forecasts are based on a simple autoregressive model with a maximum lag length of five. Given this maximum, all possible combinations of lags are considered (allow-

[18] Data are retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series.

ing "gaps") and a choice among these made using the Hannan–Quinn information criterion, with the final 20 observations excluded.[19] Although no lags are selected for non-farm business sector productivity,[20] an AR(1) forecast model is employed to allow possible dynamics,

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t. \tag{11}$$

The forecast model for the other series is selected as

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-4} + u_t. \tag{12}$$

Before turning to the forecasts, Table 6 reports the results of break inference (maximum one coefficient and one variance break) obtained by application of the two-step method of Sect. 2.2 to the full sample of data for the respective model (11) or (12), together with the unconditional mean and residual standard deviations in the implied regimes (that is, up to and subsequent to an estimated break). No coefficient break is detected in (total) non-farm business sector productivity growth, but a break is found in the residual variance at 1983Q2, which implies a substantial reduction. Previous analyses (Hansen 2001; Benati 2007) note that it can be difficult to detect a mean break in US labour productivity, possibly because any change is gradual, while a variance break at 1983Q2 is in line with many studies relating to the so-called Great Moderation for the US (McConnell and Perez-Quiros 2000; Summers 2005). Coefficient breaks are detected for the other series in Table 6, with large mean reductions; the EM confidence set in particular indicates substantial uncertainty about the coefficient break dates. A variance break for durable goods sector productivity growth is also detected, which may or may not be concurrent with the coefficient break date. Although not shown, analysis of the full sample data does not point to any break during our post-sample forecast period.[21]

To conduct the forecasts, data to 2013Q4 are initially used for estimation and testing, with a one-step ahead forecast computed for 2014Q1. Structural break tests (for coefficients and, where appropriate, residual variances) allow a single[22] break with trimming of $\epsilon = 10\%$. Data for 2014Q1 are added and forecasts for 2014Q2 are computed in the same way, and so on through the remaining period. Although repeated application of structural break tests may raise a multiple testing problem (Robbins 1970; Chu et al. 1996) , generally the same conclusions are reached as to the existence or non-existence of breaks with the estimates of break dates remaining at the same temporal locations.

[19] We also considered re-selecting models as each observation is added during the pseudo forecast period. However, the selected models remain unaltered, except that the last (fourth) lag is dropped for the manufacturing durable goods sector for some estimation samples.

[20] Akaike, Schwarz and Hannan–Quinn information criteria also selected no lags.

[21] This was confirmed by applying the Bai and Perron (1998) multiple coefficient breaks test procedure with HC inference over the full sample, allowing a maximum of 3 breaks.

[22] We also experimented by allowing a maximum of three structural breaks, but found no additional breaks except for one additional variance break in each of the manufacturing and durable goods productivity series. Allowing for these was found to make no substantive change in the forecast errors.

Table 7 shows that no method exhibits a forecast accuracy gain over the full sample benchmark model for non-farm business sector productivity growth; this is in line with the simulation results of Sect. 4 when no coefficient break occurs, as indicated in Table 6 for this series. Although not always the case in the simulations when a variance decrease applies, the two-step procedure here leads to reduced forecast errors (measured by MSFE) compared to HC inference. Averaging using either the BP confidence interval or the EM set improves accuracy over the post-break estimator, with the EM set leading to more accurate forecasts. Again in line with Sect. 4, for both types of inference the relative MSFE is reduced to close to one by use of the EM set and stepwise testing; for HC inference, the reduction is about 15% compared with the post-break method. The trade-off method also reduces MSFE compared to the post-break estimator, but is inferior to averaging over the EM set with HC inference (as in the simulations), while cross-validation loses very little compared with the benchmark model.

Since large breaks apparently occur in the coefficients of the two manufacturing sector productivity series (Table 6), it is unsurprising that all methods perform better in Table 7 in forecasting these series than the full sample benchmark model, with this especially true for methods which use information about the estimated break date. Our confidence interval/set methods perform well for both series, with averaging using the EM set yielding the smallest MSFE value across all methods. HC and two-step procedures lead to almost identical results for manufacturing productivity growth, where there is apparently no variance break, except when the stepwise testing procedure is used in combination with averaging over a confidence interval or set. The situation where stepwise testing leads to poorer forecasts than averaging alone was also noted in our simulations and can be associated with the variance decrease. For the manufacturing durable goods series, which apparently experiences both coefficient and variance reductions, two-step inference leads to more accurate forecasts than does the use of HC inference across all methods. Indeed, the improvements here from two-step inference are more impressive than indicated for DGP14 in Sect. 4. Finally, the relatively poor performance of the methods in Panel C of Table 7 is unsurprising, since the coefficient break occurs towards the end of the estimation period.

The Diebold and Mariano (1995) test is used to check whether the differences in forecast accuracy between a given forecast method and the full sample benchmark model are statistically significant. Improvements are significant at 10% or less for most methods when analysing the manufacturing sector productivity series. Two-step inference generally delivers forecasts that significantly improve on the benchmark for the manufacturing durable goods sector model, but this is not the case when HC inference is employed. No statistical evidence is found against equal predictive accuracy for all methods against the benchmark for the non-farm business sector; this is unsurprising since there is apparently no coefficient break for this case (Table 6).

Finally, to check the robustness of the results to the choice of the out-of-sample window, forecast accuracy measures are re-calculated for forecast samples of 15, 25 and 30 quarters. The results in Appendix Table 12 show that the good performance of the proposed confidence interval/set methods remains robust, yielding the smallest MSFE value for most forecast samples. Other results are also generally robust, except that (compared with Table 7) forecast accuracy relative to the benchmark deteriorates

Table 6 US productivity growth: estimated break dates and length of confidence interval/set

Estimated break dates		Length of confidence interval/set		Unconditional mean	Residual standard deviations
		BP	EM		
(a) Non-farm business sector productivity growth, output per hour worked					
Coefficient break	None			1.9	
Variance break	1983Q2	52	48		4.3; 2.4
(b) Manufacturing sector productivity growth, output per hour worked					
Coefficient break	2010Q2	19	25	3.6; 0.2	
Variance break	None				3.2
(c) Manufacturing durable goods sector productivity growth, output per hour worked					
Coefficient break	2011Q1	13	41	4.1; 0.5	
Variance break	2010Q3	60	42		4.7; 3.1

The break inference results in (a) are obtained using the full sample between 1970Q1 and 2018Q4, and the results reported in (b) and (c) use the available data from 1970Q1 to 2018Q4. None indicates no break is detected. Unconditional mean and residual standard deviations are estimated in each corresponding coefficient and variance break regime through the two-step procedure of Sect. 2.2

Table 7 US productivity growth: MSFE ratios

Forecast methods	Non-farm business sector	Manufacturing sector	Manufacturing durable goods sector
Panel A: HC break inference			
Post-break	1.172	0.864*	0.905
Trade-off	1.147	0.874**	0.893
Stepwise coefficient testing	1.156	0.850*	0.893
BP confidence interval average	1.090	0.815**	0.883
EM confidence interval average	1.052	0.775**	0.861
BP confidence interval and stepwise testing	1.056	0.806**	0.889
EM confidence interval and stepwise testing	1.040	0.790**	0.863
Panel B: Two-step break inference			
Post-break	1.017	0.864*	0.797
Trade-off	1.000	0.874**	0.809
Stepwise coefficient testing	1.000	0.845*	0.754*
BP confidence interval average	1.000	0.815**	0.792*
EM confidence interval average	1.000	0.775**	0.755**
BP confidence interval and stepwise testing	1.000	0.858*	0.768*
EM confidence interval and stepwise testing	1.000	0.818*	0.772*
Panel C: Unknown break inference			
Cross-validation (CV)	1.018	0.927	0.904
Average over all windows (CV weight)	1.068	0.973	0.901**
Average over all windows (equal weight)	1.021	0.830**	0.758***

MSFEs are reported relative to the associated MSFEs based on the full sample benchmark model using 20 quarterly out-of-sample observations. The equality of forecast accuracy of forecast methods against the full sample benchmark is tested using the Diebold and Mariano (1995) test statistic. One, two and three asterisks denote significance at 10%, 5% and 1%, respectively

in the manufacturing and manufacturing durables series when the forecast window length is 30 quarters. This relates to the coefficient break detected in each case around 2010, leaving relatively few observations for estimation of the forecast models in some sub-samples.

6 Conclusion

This paper investigates the usefulness for forecasting of employing a wider range of information relating to structural break testing than implied by the use of a point estimate of the break date in the model's coefficients. In particular, we propose using a forecast combination approach based on the confidence interval or confidence set for the estimated break date, thereby avoiding using a single and potentially poor break date estimate. In this context, we investigate the confidence interval associated with Bai and Perron (1998) and the confidence set of Eo and Morley (2015). Our simulation results show that the Eo and Morley (2015) set is particularly useful for this purpose and performs well relative to other methods, including those based on a point estimate of the break date (namely post-break and trade-off methods) and others that do not use any break information (cross-validation and averaging across all possible windows). Although testing whether breaks apply to individual coefficients can further improve forecast accuracy, it is not recommended that such a testing down procedure be used when the disturbance variance declines during the sample period.

A second issue related to inference that we examine concerns the treatment of possible breaks in the disturbance variance, comparing results based on heteroskedasticity consistent coefficient break tests with two-step inference and use of FGLS estimation. Our results show that two-step inference generally reduces forecast errors in moderate or small samples when the true coefficients are constant over time or when the variance exhibits change at the same time or subsequent to a coefficient break. Further, when two-step inference does not lead to improved forecast accuracy, its use does not involve a substantive deterioration either.

An application to US productivity growth underlines the practical usefulness of the methods proposed in the paper for forecasting in the presence of structural breaks. Our analysis considers the situation where at most one structural break applies to each of the coefficients and the disturbance variance, with the two possible breaks not necessarily concurrent; in further work, we plan to examine situations where these characteristics may each be subject to multiple breaks.

Acknowledgements The authors would like to thank Heather Anderson, Jing Tian and the editor and referees of this journal for their helpful comments on an earlier version of the paper. We also thank Ralf Becker for his many comments at all stages of this work. However, the authors take full responsibility for any errors or omissions.

Funding No funding was received for conducting this study.

Data and computer code availability The data used in the empirical section of this study are openly available in FRED, Federal Reserve Bank of St. Louis; at https://fred.stlouisfed.org, reference number (PRS85006091; PRS30006092; PRS31006092). Computer code to obtain empirical and simulation results of the paper is available from either author upon reasonable request. The code is based on codes for the

Bai and Perron (2003) and Eo and Morley (2015) procedures made available by Pierre Perron and James Morley, respectively, on their websites.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Appendix A: Other forecast methods

Cross-validation

The cross-validation approach proposed by Pesaran and Timmermann (2007) considers all possible estimation windows of different lengths and chooses the single window which achieves the smallest pseudo out of sample forecast error. Specifically, for each possible starting point m for the estimation window, the one which generates the smallest MSFE is selected as

$$m^*(T, \tilde{w}, \underline{w}) = \arg \min_{m=1,\ldots,T-\tilde{w}-\underline{w}} \left\{ \tilde{w}^{-1} \sum_{t=T-\tilde{w}}^{T-1} (y_{t+1} - \mathbf{x}_t' \hat{\beta}_{m:t})^2 \right\}$$

where $\hat{\beta}_{m:t}$ is the OLS estimate based on the observation window $[m : t]$ and $m \in 1, \ldots, T - \tilde{w} - \underline{w}$, having a minimum estimation window $\underline{w}$ and reserving the last $\tilde{w}$ observations for the pseudo out of sample evaluation. The forecast model uses $\hat{\beta}'_{m^*:T}$ estimated over the sample $[m^* : T]$. We use $\underline{w}$=0.1T and $\tilde{w}$=0.25T.

Trade-off

This method trades off bias against forecast error variance by selecting v_1 to minimize (Pesaran and Timmermann 2007)

$$f(v_1) = \lambda^2(\mu' \sum\nolimits_{v_1} \sum\nolimits_{v}^{-1} \mathbf{x}_T)^2 + \frac{1}{v}(\mathbf{x}_T' \sum\nolimits_{v}^{-1} \mathbf{x}_T)^2 + \frac{\lambda\psi}{v}(\mathbf{x}_T' \sum\nolimits_{v}^{-1} \sum\nolimits_{v_1} \sum\nolimits_{v}^{-1} \mathbf{x}_T)$$

where $\mu = (\hat{\beta}_2 - \hat{\beta}_1)/\hat{\sigma}_2$, $\psi = (\hat{\sigma}_1^2 - \hat{\sigma}_2^2)/\hat{\sigma}_2^2$, $\lambda = v_1/v$ and $v = v_1 + v_2$, with v_1 and v_2 the number of pre- and post-break observations, respectively, $\hat{\beta}_i$ and $\hat{\sigma}_i^2$ $(i = 1, 2)$

the respective coefficient and variance estimates, and

$$\sum\nolimits_{v_1} = v_1^{-1} \sum_{t=m}^{\hat{T}_c} \mathbf{x}'_{t-1}\mathbf{x}_{t-1}, \quad \sum\nolimits_{v_2} = v_2^{-1} \sum_{\hat{T}_c+1}^{T} \mathbf{x}'_{t-1}\mathbf{x}_{t-1}, \quad \sum\nolimits_{v}$$
$$= \lambda \sum\nolimits_{v_1} + (1-\lambda) \sum\nolimits_{v_2}.$$

With v_1 selected to minimize this function, the coefficient vector used in forecasting is $\hat{\beta}_{\hat{T}_c - v_1 + 1:T}$ estimated over the sample of $[\hat{T}_c - v_1 + 1 : T]$.

Forecast combination over estimation samples

Forecasts using the same model estimated over different sizes of windows are averaged to generate a single forecast for $T+1$ (Pesaran and Timmermann 2007; Pesaran and Pick 2011). Specifically, in our notation and using equal weights,

$$\hat{y}_{T+1}(T, \underline{w}) = (T - \underline{w})^{-1} \sum_{m=1}^{T-\underline{w}} (\mathbf{x}'_T \hat{\beta}_{m:T}).$$

Using weights proportional to the inverse of the associated pseudo out of sample MSFE values (Pesaran and Timmermann 2007), the cross-validation weighted average is

$$\hat{y}_{T+1}(T, \tilde{w}, \underline{w}) = \frac{\sum_{m=1}^{T-\tilde{w}-\underline{w}} (\mathbf{x}'_T \hat{\beta}_{m:T}) \, [MSFE(m|T, \tilde{w})]^{-1}}{\sum_{m=1}^{T-\tilde{w}-\underline{w}} [MSFE(m|T, \tilde{w})]^{-1}}.$$

Appendix B: MSFE ratios, no disturbance in the forecast period

See Table 8.

Appendix C: Larger sample performance tables

See Table 9.

Table 8 MSFE ratios, concurrent mid-sample breaks, $T = 100$, $\varepsilon_{T+1} = 0$

	Coefficients constant			Overall mean constant							Dynamics constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
Panel A: HC break inference														
Post-break	4.191	8.185	1.215	0.117	0.052	0.407	0.073	0.078	2.318	0.020	1.119	0.159	3.446	0.043
Trade-off	2.995	5.609	1.087	0.127	0.096	0.369	0.077	0.084	1.719	0.035	0.910	0.155	2.412	0.056
Stepwise testing	3.977	7.986	1.202	0.106	0.046	0.423	0.069	0.079	2.363	0.021	1.001	0.142	3.215	0.055
BP confidence interval average	2.780	4.785	1.085	0.141	0.081	0.365	0.074	0.089	1.425	0.021	0.864	0.152	2.022	0.042
EM confidence interval average	1.751	1.746	1.027	0.112	0.053	0.335	0.071	0.077	0.927	0.020	0.790	0.169	1.117	0.042
BP confidence interval and stepwise testing	2.664	4.412	1.062	0.126	0.062	0.379	0.072	0.088	1.538	0.021	0.773	0.133	1.867	0.051
EM confidence interval and stepwise testing	1.606	1.522	0.993	0.105	0.053	0.339	0.069	0.080	0.993	0.021	0.698	0.143	0.997	0.048
Panel B: Two-step break inference														
Post-break	2.224	4.295	1.216	0.120	0.052	0.405	0.073	0.078	1.985	0.020	0.988	0.161	2.689	0.039
Trade-off	1.727	3.253	1.252	0.131	0.096	0.370	0.077	0.085	1.297	0.030	0.842	0.157	1.729	0.048
Stepwise testing	2.143	4.601	1.391	0.110	0.046	0.419	0.069	0.080	1.921	0.020	0.890	0.143	2.215	0.039

Table 8 continued

	Coefficients constant			Overall mean constant							Dynamics constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
BP confidence interval average	1.679	2.835	1.275	0.145	0.082	0.368	0.074	0.090	1.235	0.020	0.817	0.154	1.564	0.039
EM confidence interval average	1.488	1.521	1.042	0.117	0.053	0.339	0.071	0.077	0.899	0.020	0.798	0.171	1.122	0.042
BP confidence interval and stepwise testing	1.653	2.838	1.245	0.127	0.060	0.379	0.072	0.088	1.269	0.020	0.726	0.135	1.328	0.042
EM confidence interval and stepwise testing	1.391	1.467	1.009	0.107	0.051	0.341	0.069	0.081	0.992	0.020	0.696	0.144	0.901	0.044
Panel C: Unknown break inference														
Cross-validation (CV)	2.003	2.654	1.180	0.147	0.076	0.341	0.093	0.105	1.002	0.026	0.695	0.227	1.392	0.057
Average over all windows (CV weight)	1.428	2.071	0.809	0.287	0.257	0.358	0.200	0.225	0.624	0.101	0.509	0.323	0.866	0.174
Average over all windows (equal weight)	1.965	2.922	0.955	0.236	0.226	0.335	0.193	0.199	0.733	0.161	0.516	0.269	1.037	0.190

The details of the DGPs are presented in Table 1. T is the total sample size and the true break date is in the middle of the sample. MSFEs are reported relative to the associated MSFE based on the full sample benchmark model. The forecast period disturbance is set to $\varepsilon_{T+1} = 0$ in (1)

Table 9 MSFE ratios, larger sample performance of concurrent breaks, $T = 200$

Forecast methods/DGPs	Coefficients constant			Overall mean constant							Dynamics constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean break	Large Mean break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
Panel A: HC break inference														
Post-break	1.032	1.061	1.006	0.535	0.411	0.774	0.453	0.465	0.956	0.184	0.916	0.648	1.045	0.313
Trade-off	1.023	1.041	1.004	0.542	0.429	0.774	0.454	0.468	0.947	0.189	0.912	0.648	1.024	0.317
Stepwise coefficient testing	1.033	1.058	1.006	0.533	0.410	0.772	0.451	0.465	0.959	0.184	0.911	0.644	1.039	0.314
BP confidence interval average	1.011	1.021	1.001	0.548	0.424	0.775	0.453	0.470	0.941	0.184	0.909	0.648	1.001	0.313
EM confidence interval average	1.011	1.003	0.998	0.535	0.412	0.773	0.453	0.465	0.935	0.184	0.914	0.648	0.987	0.314
BP confidence interval and stepwise testing	1.011	1.015	1.002	0.541	0.415	0.772	0.451	0.468	0.944	0.184	0.902	0.643	0.994	0.314
EM confidence interval and stepwise testing	1.011	1.003	0.999	0.533	0.411	0.771	0.451	0.465	0.937	0.184	0.907	0.643	0.981	0.315
Panel B: Two-step break inference														
Post-break	1.019	1.031	1.004	0.535	0.411	0.773	0.452	0.465	0.948	0.184	0.917	0.648	1.011	0.313
Trade-off	1.014	1.026	1.007	0.542	0.429	0.774	0.454	0.468	0.944	0.187	0.913	0.648	0.993	0.315
Stepwise coefficient testing	1.018	1.037	1.011	0.533	0.410	0.772	0.451	0.465	0.951	0.184	0.913	0.644	0.999	0.312

Table 9 continued

Forecast methods/DGPs	Coefficients constant			Overall mean constant							Dynamics constant			
	No break	σ x 4	σ x 0.5	Small Break in β_{11}	Large Break in β_{11}	Small Break in β_{12}	Large Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean break	Large Mean break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
BP confidence interval average	1.009	1.017	1.004	0.547	0.424	0.775	0.453	0.470	0.988	0.184	0.909	0.648	1.006	0.313
EM confidence interval average	1.009	1.007	1.000	0.535	0.412	0.773	0.452	0.465	0.950	0.184	0.914	0.648	0.989	0.314
BP confidence interval and stepwise testing	1.008	1.015	1.003	0.541	0.415	0.773	0.451	0.468	0.980	0.184	0.904	0.643	0.993	0.313
EM confidence interval and stepwise testing	1.009	1.006	1.000	0.533	0.411	0.771	0.451	0.465	0.949	0.184	0.908	0.643	0.976	0.313
Panel C: Unknown break inference														
Cross-validation (CV)	1.017	1.015	1.008	0.541	0.416	0.783	0.459	0.471	0.940	0.186	0.921	0.660	1.000	0.318
Average over all windows (CV weight)	1.008	1.011	0.994	0.643	0.546	0.819	0.534	0.562	0.940	0.251	0.918	0.733	0.979	0.412
Average over all windows (equal weight)	1.016	1.018	0.999	0.619	0.531	0.804	0.529	0.550	0.936	0.310	0.912	0.708	0.980	0.427

The details of the DGPs are presented in Table 1. T is the total sample size and the true break date is in the middle of the sample. MSFEs are reported relative to the associated MSFE based on the full sample benchmark model

Table 10 MSFE ratios, larger sample performance of non-concurrent breaks, $T = 200$

Forecast methods/DGPs	Overall mean constant					Dynamics constant			
	Small Break in β_{11}	Small Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 4	DGP 6	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
Tc= 0.25T, Tv = 0.50T, T = 200									
Panel A: HC break inference									
Post-break	0.650	0.916	0.659	0.995	0.343	0.984	0.780	1.058	0.479
Trade-off	0.656	0.917	0.662	0.984	0.347	0.980	0.781	1.035	0.481
EM confidence interval average	0.650	0.916	0.659	0.971	0.342	0.971	0.781	0.997	0.479
EM confidence interval & stepwise testing	0.650	0.915	0.659	0.974	0.342	0.968	0.779	0.997	0.481
Panel B: Two-step break inference									
Post-break	0.650	0.916	0.659	0.953	0.341	0.983	0.780	0.979	0.477
Trade-off	0.655	0.916	0.661	0.954	0.342	0.979	0.781	0.978	0.477
EM confidence interval average	0.650	0.916	0.659	0.958	0.341	0.971	0.780	0.983	0.479
EM confidence interval & stepwise testing	0.650	0.915	0.659	0.961	0.341	0.968	0.780	0.987	0.479
Panel C: Unknown break inference									
Cross-validation (CV)	0.662	0.932	0.673	0.979	0.347	0.976	0.796	0.998	0.488
Average over all windows (CV weight)	0.671	0.920	0.678	0.968	0.354	0.964	0.795	0.988	0.492
Average over all windows (equal weight)	0.673	0.924	0.680	0.974	0.367	0.970	0.796	0.995	0.497

Table 10 continued

Forecast methods/DGPs	Overall mean constant					Dynamics constant			
	Small Break in β_{11}	Small Break in β_{12}	Break in β_{11} β_{12}	DGP 8 σ x 4	DGP 8 σ x 0.5	Small Mean Break	Large Mean Break	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 4	DGP 6	DGP 8	DGP 9	DGP 10	DGP 11	DGP 12	DGP 13	DGP 14
Tc = 0.75T, Tv = 0.50T, T = 200									
Panel A: HC break inference									
Post-break	0.479	0.676	0.379	0.964	0.135	0.896	0.515	1.063	0.201
Trade-off	0.486	0.673	0.383	0.946	0.141	0.883	0.517	1.034	0.205
EM confidence interval average	0.478	0.673	0.380	0.956	0.136	0.878	0.521	0.991	0.205
EM confidence interval & stepwise testing	0.477	0.667	0.380	0.963	0.136	0.856	0.510	0.986	0.202
Panel B: Two-step break inference									
Post-break	0.479	0.677	0.379	0.989	0.135	0.899	0.515	1.058	0.200
Trade-off	0.487	0.673	0.383	0.958	0.141	0.886	0.517	1.017	0.205
EM confidence interval average	0.478	0.673	0.380	0.972	0.136	0.882	0.521	0.991	0.205
EM confidence interval & stepwise testing	0.477	0.668	0.381	0.975	0.136	0.858	0.510	0.983	0.202
Panel C: Unknown break inference									
Cross-validation (CV)	0.704	0.763	0.587	0.942	0.411	0.900	0.698	0.994	0.485
Average over all windows (CV weight)	0.851	0.866	0.812	0.968	0.727	0.927	0.795	0.988	0.645
Average over all windows (equal weight)	0.702	0.773	0.638	0.929	0.500	0.879	0.681	0.974	0.483

The details of the DGPs are presented in Table 1. T is the total sample size, T_c is the location of true coefficient break and T_v is the location of true disturbance variance break

Table 11 MSFE ratios, late variance break, $T = 200$

Forecast methods/DGPs	Coefficients constant		Overall mean constant		Dynamics constant	
	σ x 4	σ x 0.5	DGP 8 σ x 4	DGP 8 σ x 0.5	DGP 12 σ x 4	DGP 12 σ x 0.5
	DGP 2	DGP 3	DGP 9	DGP 10	DGP 13	DGP 14
Tc = 0.50T, Tv = 0.75T, T= 200						
Panel A: HC break inference						
Post-break	1.064	1.013	0.952	0.193	1.047	0.326
Trade-off	1.043	1.009	0.940	0.197	1.024	0.326
EM confidence interval average	1.007	1.004	0.926	0.192	0.993	0.325
EM confidence interval and stepwise testing	1.007	1.001	0.931	0.192	0.991	0.322
Panel B: Two-step break inference						
Post-break	1.033	0.996	0.896	0.189	0.956	0.320
Trade-off	1.024	1.004	0.897	0.190	0.956	0.320
EM confidence interval average	1.013	1.000	0.904	0.189	0.968	0.320
EM confidence interval and stepwise testing	1.012	0.999	0.909	0.189	0.970	0.318
Panel C: Unknown break inference						
Cross-validation (CV)	1.014	1.011	0.920	0.192	0.989	0.328
Average over all windows (CV weight)	1.008	1.004	0.932	0.261	0.978	0.422
Average over all windows (equal weight)	1.018	1.000	0.927	0.311	0.980	0.427

The details of the DGPs are presented in Table 1. T is the total sample size, T_c is the location of true coefficient break and T_v is the location of true disturbance variance break

Table 12 MSFE ratios with different out-of-sample windows

Forecast methods	Non-farm business sector			Manufacturing sector			Manufacturing durable goods sector		
	Q=15	Q=25	Q=30	Q=15	Q=25	Q=30	Q=15	Q=25	Q=30
Panel A: HC break inference									
Post-break	1.007	1.102	1.008	0.863	0.858*	0.883**	1.015	0.912	0.983
Trade-off	1.002	1.079	1.006	0.853*	0.825***	0.887***	1.009	0.910	0.994
Stepwise coefficient testing	1.007	1.102	1.008	0.841	0.786**	0.830***	1.002	0.895	0.990
BP confidence interval average	0.996	1.050	1.004	0.801*	0.818**	0.846***	0.965	0.889	0.945
EM confidence interval average	0.996	1.018	1.000	0.805*	0.797**	0.821***	0.947	0.869*	0.923**
BP confidence interval and stepwise testing	0.997	1.033	1.002	0.795*	0.781**	0.816***	0.968	0.891	0.954
EM confidence interval and stepwise testing	0.997	1.011	1.000	0.822*	0.791**	0.813***	0.948	0.870*	0.928**
Panel B: Two-step break inference									
Post-break	1.021	1.004	1.009	0.867	0.858*	0.883**	0.831	0.834	0.842*
Trade-off	1.000	1.000	1.000	0.859*	0.825***	0.887***	0.897	0.833	0.864*
Stepwise coefficient testing	1.000	1.000	1.000	0.816	0.834**	0.854***	0.751*	0.792	0.812*
BP confidence interval average	1.000	1.000	1.000	0.788*	0.818**	0.846***	0.816*	0.823*	0.825**
EM confidence interval average	1.000	1.000	1.000	0.798*	0.797**	0.821***	0.712**	0.791**	0.790***
BP confidence interval & stepwise testing	1.000	1.000	1.000	0.853	0.817**	0.844***	0.781*	0.806*	0.830**
EM confidence interval & stepwise testing	1.000	1.000	1.000	0.849	0.802**	0.830***	0.733*	0.804**	0.814**
Panel C: Unknown break inference									
Cross-validation (CV)	0.978	1.002	0.995	1.003	0.890	0.874**	0.878	0.894	0.906
Average over all windows (CV weight)	1.050	1.047	1.047	1.018	0.959	0.970	0.934	0.883**	0.903**
Average over all windows (equal weight)	0.965**	1.016	1.014	0.864*	0.816***	0.846***	0.790**	0.736***	0.788***

MSFEs are reported relative to the associated MSFEs based on the full sample benchmark model. The equality of forecast accuracy of forecast methods against the full sample benchmark is tested using the Diebold and Mariano (1995) test statistic. One, two and three asterisks denote significance at 10%, 5% and 1%, respectively. Q indicates the length of out-of-sample forecast values in quarters

References

Altansukh G, Becker R, Bratsiotis G, Osborn DR (2017) What is the globalisation of inflation? J Econ Dyn Control 74:1–27

Bai J (1997) Estimation of a change point in multiple regression models. Rev Econ Stat 79:551–563

Bai J, Perron P (1998) Estimating and testing linear models with multiple structural changes. Econometrica 66:47–78

Bai J, Perron P (2006) Multiple structural change models: a simulation analysis. In: Corbea D, Durlauf S, Hansen BE (eds) Econometric theory and practice: frontiers of analysis and applied research. Cambridge University Press, Cambridge, pp 212–237

Bataa E, Osborn DR, Sensier M, van Dijk D (2013) Structural breaks in the international dynamics of inflation. Rev Econ Stat 95:646–659

Benati L (2007) Drift and breaks in labor productivity. J Econ Dyn Control 31:2847–2877

Boot T, Pick A (2020) Does modeling a structural break improve forecast accuracy? J Econom 215:35–59

Chang SY, Perron P (2018) A comparison of alternative methods to construct confidence intervals for the estimate of a break date in linear regression models. Econom Rev 37:577–601. https://doi.org/10.1080/07474938.2015.1122142

Chu C-SJ, Stinchcombe M, White H (1996) Monitoring structural change. Econometrica 64:1045–1065

Clark TE, McCracken MW (2005) The power of tests of predictive ability in the presence of structural breaks. J Econom 124:1–31

Davidian M, Carroll RJ (1987) Variance function estimation. J Am Stat Assoc 82:1079–1091

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13:253–263

Eklund J, Kapetanios G, Price S (2013) Robust forecast methods and monitoring during structural change. Manch Sch 81:3–27

Elliott G (2005) Forecasting when there is a single break. Manuscript. University of California, San Diego

Elliott G, Müller UK (2007) Confidence sets for the date of a single break in linear time series regressions. J Econom 141:1196–1218

Eo Y, Morley J (2015) Likelihood-ratio-based confidence sets for the timing of structural breaks: likelihood-ratio-based confidence sets. Quant Econ 6:463–497

Hansen BE (2001) The new econometrics of structural change: dating breaks in U.S. labor productivity. J Econ Perspect 15:117–128

Hendry DF (2000) On detectable and non-detectable structural change. Struct Chang Econ Dyn 11:45–65

Hendry DF, Clements MP (2003) Economic forecasting: some lessons from recent research. Econ Model 20:301–329

Inoue A, Jin L, Rossi B (2017) Rolling window selection for out-of-sample forecasting with time-varying parameters. J Econom 196:55–67

Jorgenson DW, Ho MS, Stiroh KJ (2008) A retrospective look at the U.S. productivity growth resurgence. J Econ Perspect 22:3–24

Koo B, Seo MH (2015) Structural-break models under mis-specification: implications for forecasting. J Econom 188:166–181

McConnell MM, Perez-Quiros G (2000) Output fluctuations in the united states: what has changed since the early 1980s? Am Econ Rev 90:1464–1476

Paye BS, Timmermann A (2006) Instability of return prediction models. J Empir Financ 13:274–315

Pesaran MH, Pick A (2011) Forecast combination across estimation windows. J Bus Econ Stat 29:307–318

Pesaran MH, Timmermann A (2004) How costly is it to ignore breaks when forecasting the direction of a time series? Int J Forecast 20:411–425

Pesaran MH, Timmermann A (2005) Small sample properties of forecasts from autoregressive models under structural breaks. J Econom 129:183–217

Pesaran MH, Timmermann A (2007) Selection of estimation window in the presence of breaks. J Econom 137:134–161

Pesaran MH, Pick A, Pranovich M (2013) Optimal forecasts in the presence of structural breaks. J Econom 177:134–152

Pitarakis J-Y (2004) Least squares estimation and tests of breaks in mean and variance under misspecification. Econom J 7:32–54

Robbins H (1970) Statistical methods related to the law of the iterated logarithm. Ann Math Stat 41:1397–1409

Stock JH, Watson MW (1996) Evidence on structural instability in macroeconomic time series relations. J Bus Econ Stat 14:11–30
Summers PM (2005) What caused the great moderation? Some cross-country evidence. Econ Rev Federal Reserve Bank Kansas City 90:5–32
Syverson C (2017) Challenges to mismeasurement explanations for the US productivity slowdown. J Econ Perspect 31:165–186
Tian J, Anderson HM (2014) Forecast combinations under structural break uncertainty. Int J Forecast 30:161–175